Bayesian regression with a categorical predictor

BAYESIAN MODELING WITH RJAGS

Alicia Johnson
Associate Professor, Macalester College

Chapter 4 goals

- Incorporate *categorical* predictors into Bayesian models
- Engineer *multivariate* Bayesian regression models
- Extend our methodology for Normal regression models to generalized linear models: Poisson regression

Rail-trail volume

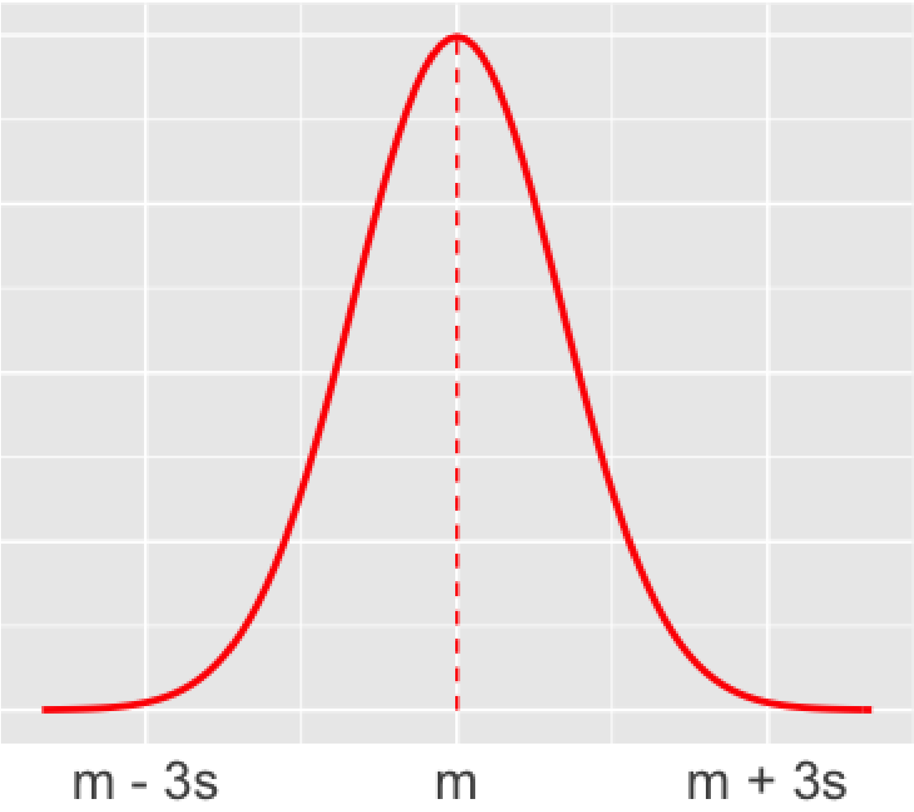
Goal:

Explore daily volume on a rail-trail in Massachusetts.

[1] Photo courtesy commons.wikimedia.org

Modeling volume

Y_i = trail volume (# of users) on day i



Model

$Y_i \sim N(m_i, s^2)$

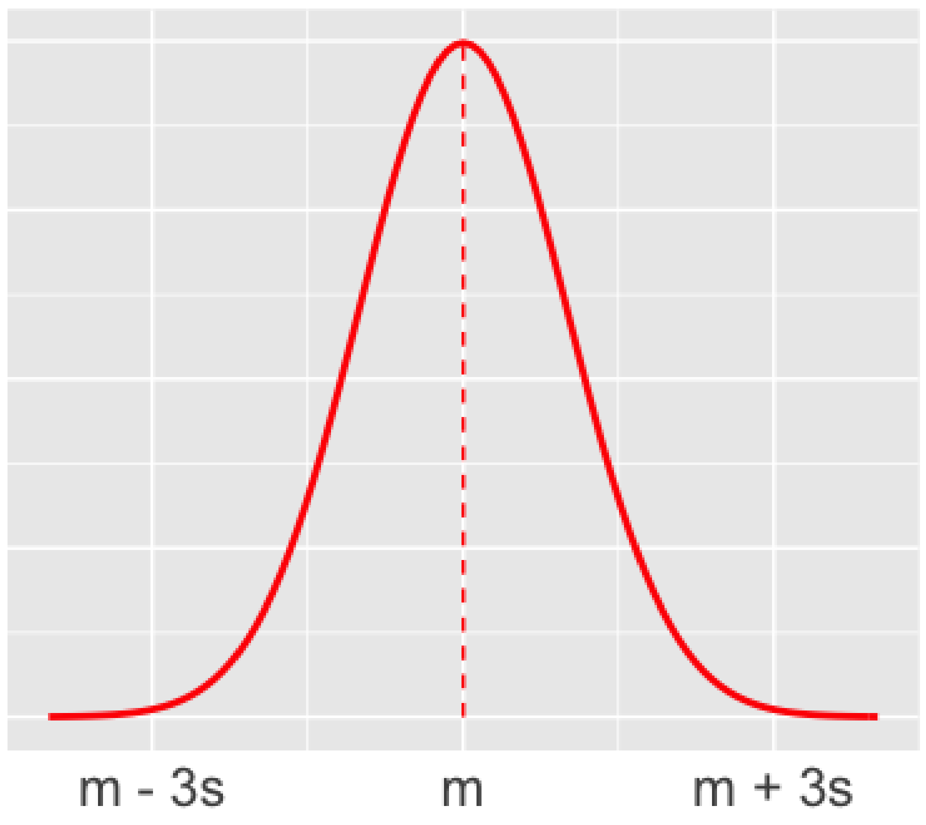
Modeling volume by weekday

Y_i = trail volume (# of users) on day i

X_i = 1 for weekdays, 0 for weekends

Model

$Y_i \sim N(m_i, s^2)$



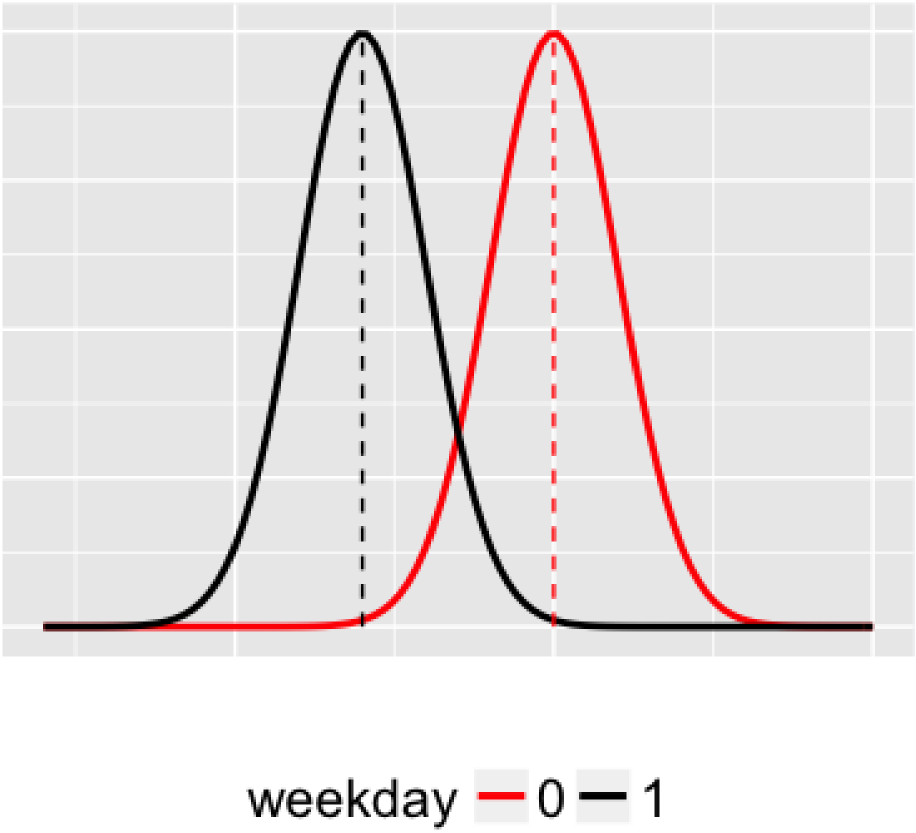
Modeling volume by weekday

Y_i = trail volume (# of users) on day i

X_i = 1 for weekdays, 0 for weekends

Model

$Y_i \sim N(m_i, s^2)$



Modeling volume by weekday

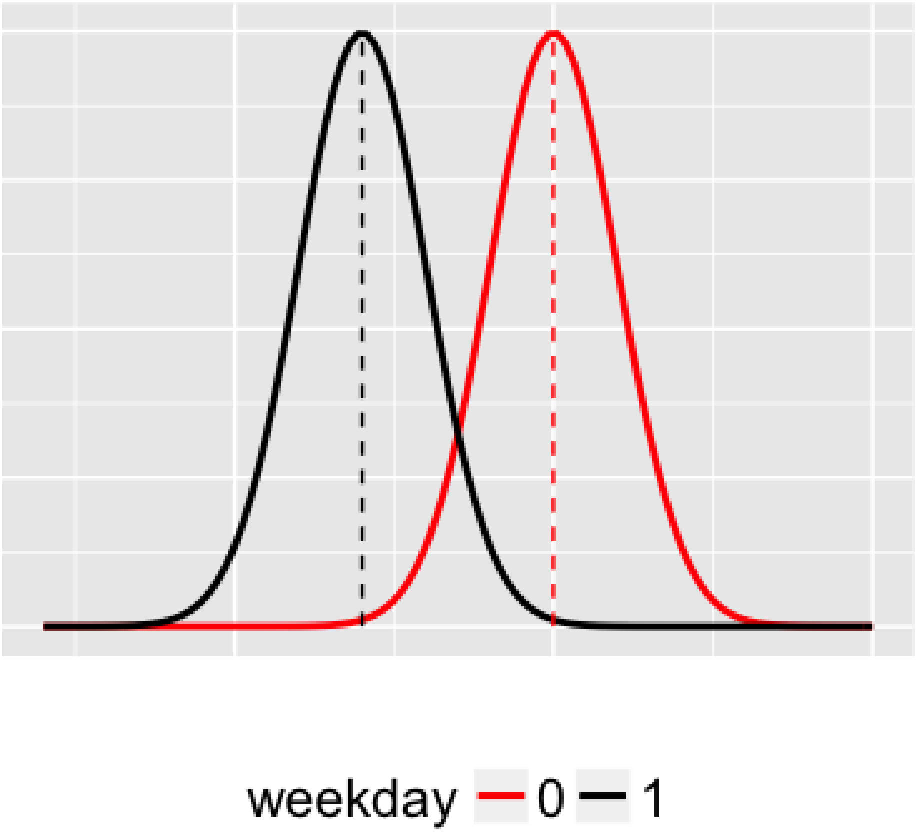
Y_i = trail volume (# of users) on day i

X_i = 1 for weekdays, 0 for weekends

Model

$Y_i \sim N(m_i, s^2)$

$m_i = a + bX_i$



Modeling volume by weekday

Y_i = trail volume (# of users) on day i

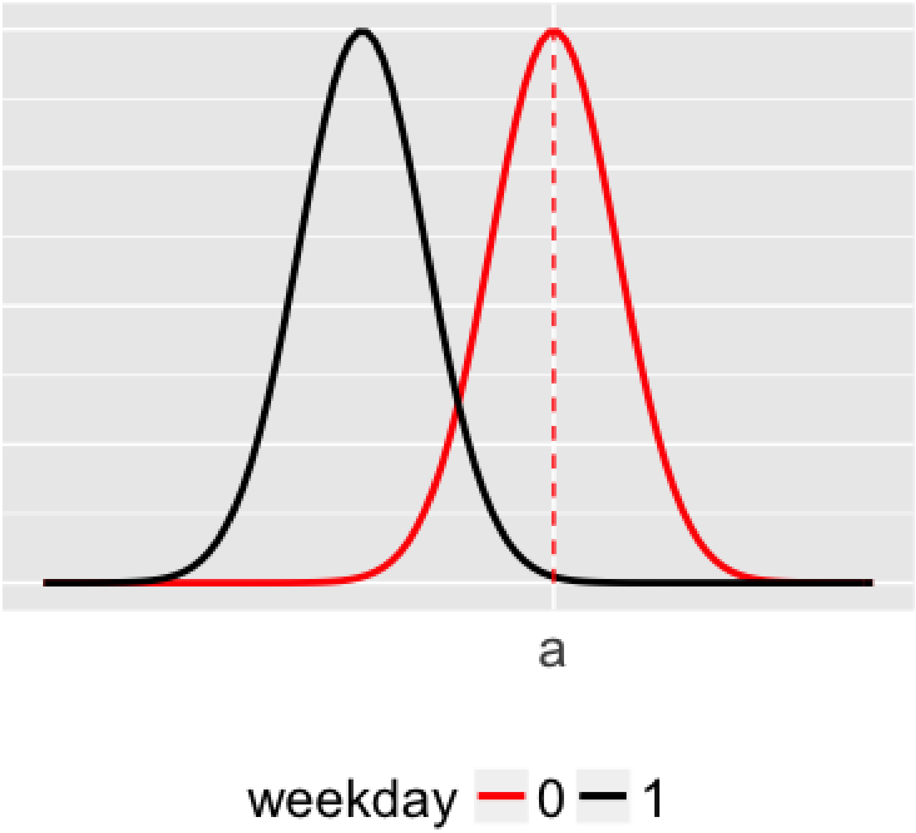
X_i = 1 for weekdays, 0 for weekends

Model

$Y_i \sim N(m_i, s^2)$

$m_i = a + bX_i$

- a = typical weekend volume



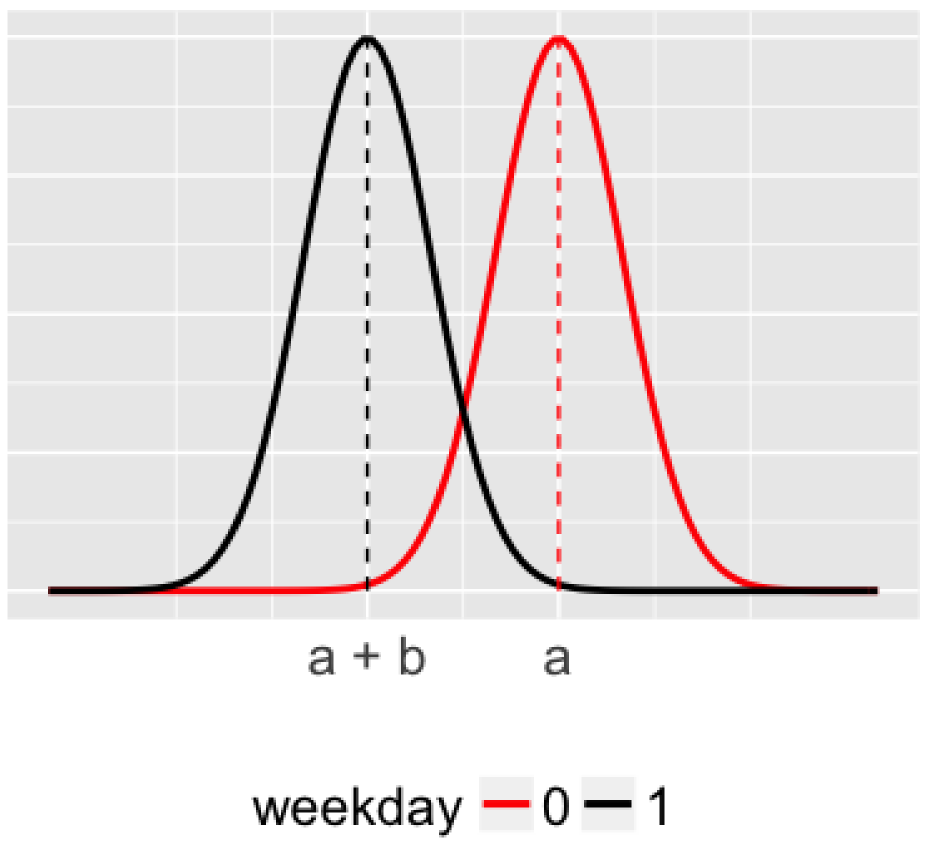
Y_i = trail volume (# of users) on day i

X_i = 1 for weekdays, 0 for weekends

Model

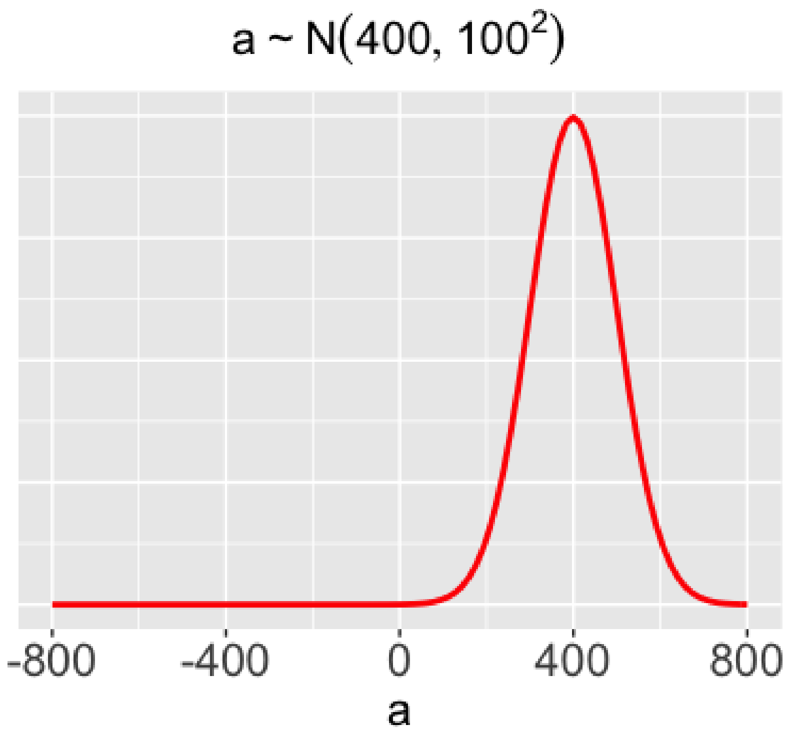
$$Y_i \sim N(m_i, s^2)$$
$$m_i = a + bX_i$$

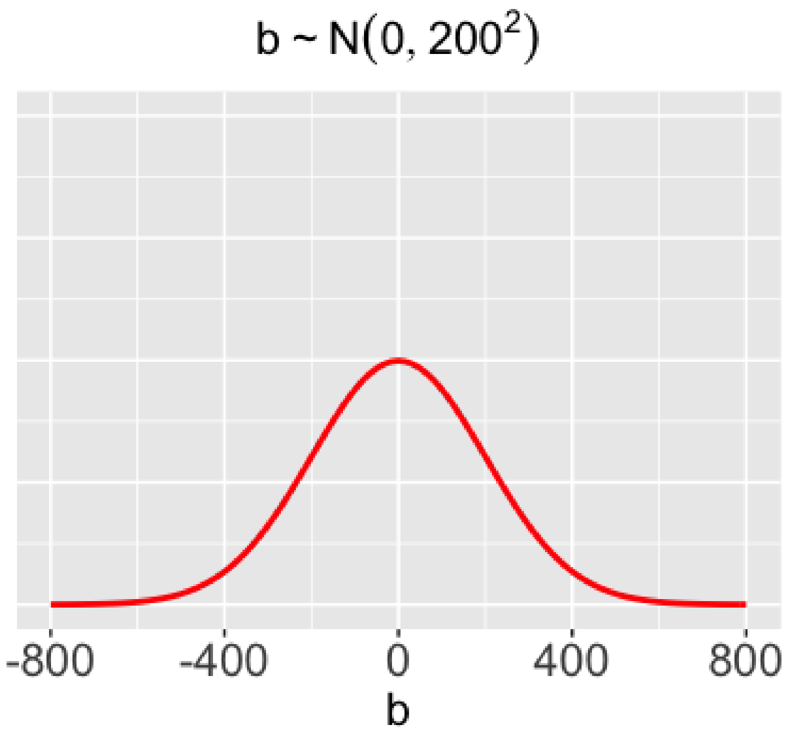
- a = typical weekend volume
- $a + b$ = typical weekday volume



- b = contrast between typical weekday vs weekend volume
- s = residual standard deviation

Priors for a & b

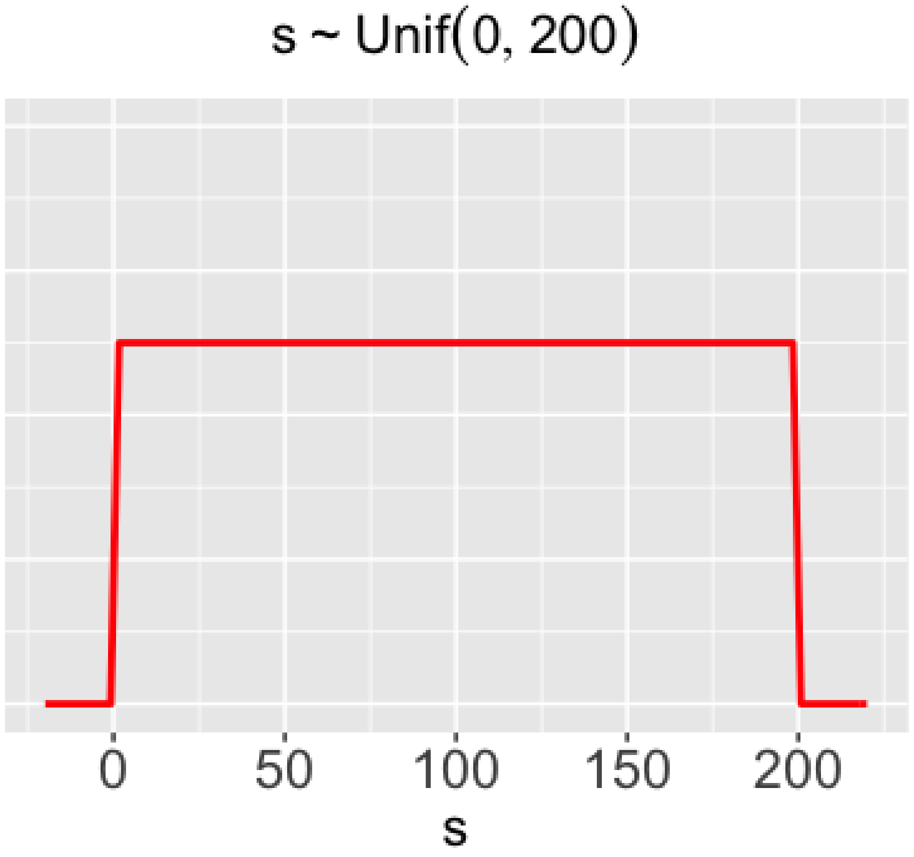




Typical *weekend* volume is most likely around 400 users per day, but possibly as low as 100 or as high as 700 users.

We lack certainty about how weekday volume compares to weekend volume. It could be more, it could be less.

Prior for s



The standard deviation in volume from day to day (whether on weekdays or weekends) is equally likely to be anywhere between 0 and 200 users.

Bayesian model of volume by weekday status

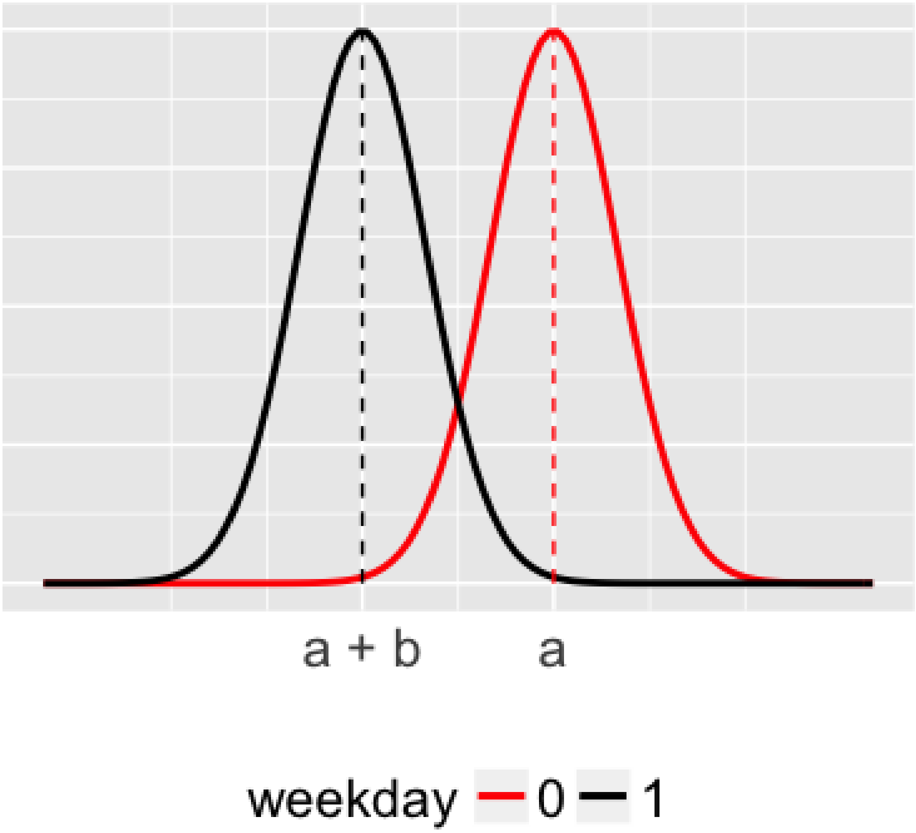
$Y_i \sim N(m_i, s^2)$

$m_i = a + bX_i$

$a \sim N(400, 100^2)$

$b \sim N(0, 200^2)$

$s \sim \text{Unif}(0, 200)$



DEFINE the Bayesian model in RJAGS

$$Y_i \sim N(m_i, s^2)$$
$$m_i = a + bX_i$$
$$a \sim N(400, 100^2)$$
$$b \sim N(0, 200^2)$$
$$s \sim \text{Unif}(0, 200)$$

```
rail_model_1 <- "model{
    # Likelihood model for Y[i]






    # Prior models for a, b, s




}"
```

DEFINE the Bayesian model in RJAGS

$$Y_i \sim N(m_i, s^2)$$
$$m_i = a + bX_i$$
$$a \sim N(400, 100^2)$$
$$b \sim N(0, 200^2)$$
$$s \sim \text{Unif}(0, 200)$$

```
rail_model_1 <- "model{
    # Likelihood model for Y[i]
    for(i in 1:length(Y)) {
        Y[i] ~ dnorm(m[i], s^(-2))

    }

    # Prior models for a, b, s
    a ~ dnorm(400, 100^(-2))
    s ~ dunif(0, 200)


}"
```

DEFINE the Bayesian model in RJAGS

`m[i] <- a + b[X[i]]`

- `X[1]` = weekend, `X[2]` = weekday
- `b` has 2 levels: `b[1]`, `b[2]`
- weekend trend ($m_i = a$) `m[i] <- a + b[1]`

```
rail_model_1 <- "model{
    # Likelihood model for Y[i]
    for(i in 1:length(Y)) {
        Y[i] ~ dnorm(m[i], s^(-2))
        m[i] <- a + b[X[i]]
    }

    # Prior models for a, b, s
    a ~ dnorm(400, 100^(-2))
    s ~ dunif(0, 200)


}"
```

DEFINE the Bayesian model in RJAGS

`m[i] <- a + b[X[i]]`

- `X[1]` = weekend, `X[2]` = weekday
- `b` has 2 levels: `b[1]`, `b[2]`
- weekend trend $(m_i = a)$
  `m[i] <- a + b[1]`
  `b[1] <- 0`

```
rail_model_1 <- "model{
    # Likelihood model for Y[i]
    for(i in 1:length(Y)) {
        Y[i] ~ dnorm(m[i], s^(-2))
        m[i] <- a + b[X[i]]
    }

    # Prior models for a, b, s
    a ~ dnorm(400, 100^(-2))
    s ~ dunif(0, 200)
    b[1] <- 0

}"
```

DEFINE the Bayesian model in RJAGS

`m[i] <- a + b[X[i]]`

- `X[1]` = weekend, `X[2]` = weekday
- `b` has 2 levels: `b[1]`, `b[2]`
- weekend trend ($m_i = a$)
  `m[i] <- a + b[1]`
  `b[1] <- 0`
- weekday ($m_i = a + b$)
  `m[i] <- a + b[2]`

```
rail_model_1 <- "model{
    # Likelihood model for Y[i]
    for(i in 1:length(Y)) {
        Y[i] ~ dnorm(m[i], s^(-2))
        m[i] <- a + b[X[i]]
    }

    # Prior models for a, b, s
    a ~ dnorm(400, 100^(-2))
    s ~ dunif(0, 200)
    b[1] <- 0
    b[2] ~ dnorm(0, 200^(-2))
}"
```

`b[2] ~ dnorm(0, 200^(-2))`

Let's practice!

BAYESIAN MODELING WITH RJAGS

Multivariate Bayesian regression

BAYESIAN MODELING WITH RJAGS

Alicia Johnson
Associate Professor, Macalester College

Modeling volume

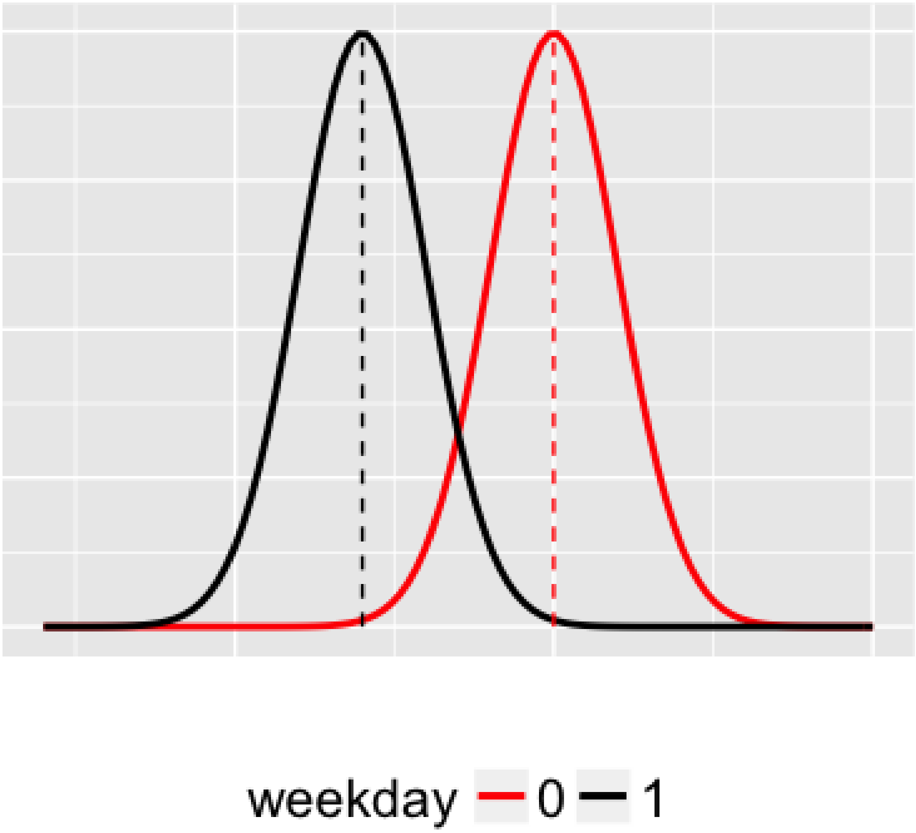
Y_i = trail volume (# of users) on day i

[1] Photo courtesy commons.wikimedia.org

Modeling volume by weekday

Y_i = trail volume (# of users) on day i

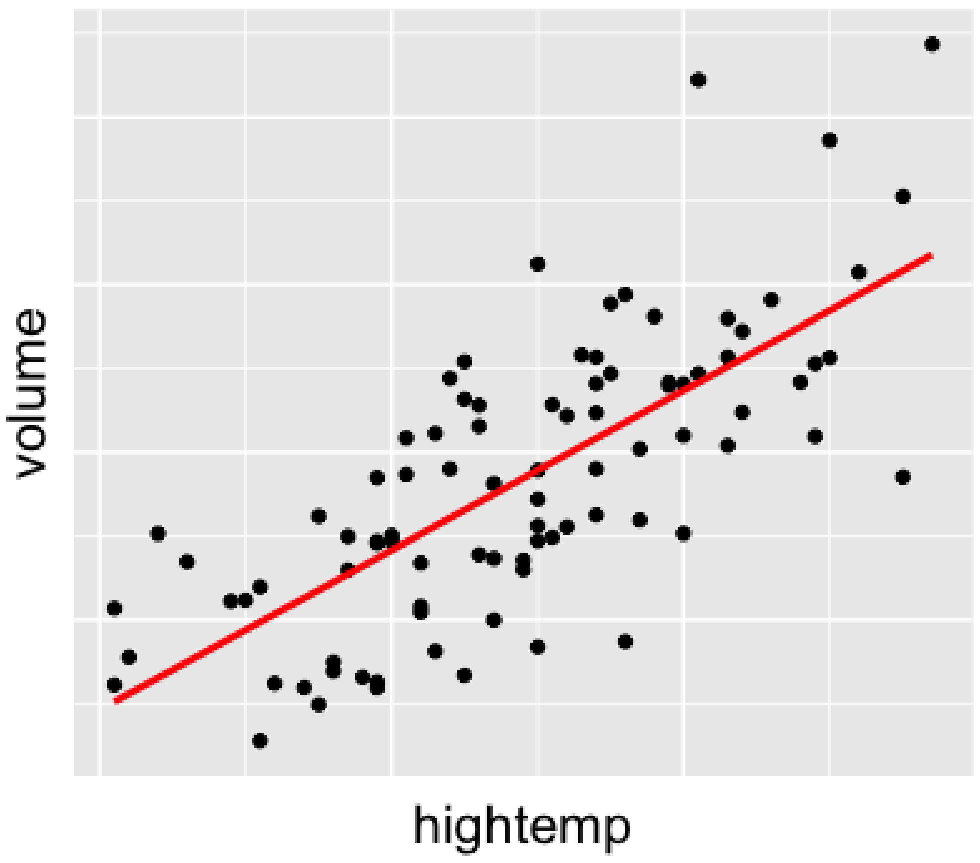
X_i = 1 for weekdays, 0 for weekends



Modeling volume by temperature

Y_i = trail volume (# of users) on day i

Z_i = high temperature on day i (in $^\circ$F)

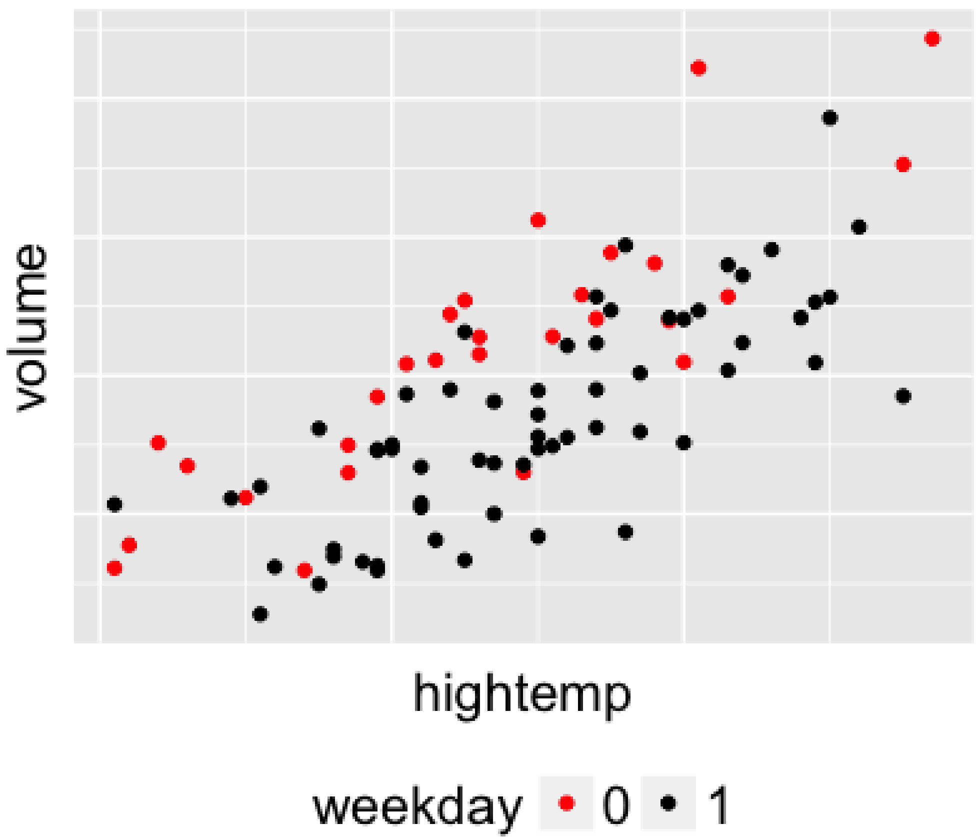


Modeling volume by temperature & weekday

Y_i = trail volume (# of users) on day i

X_i = 1 for weekdays, 0 for weekends

Z_i = high temperature on day i (in $^\circ$F)



$$Y_i \sim N(m_i, s^2)$$

$$m_i = a + bX_i + cZ_i$$

Weekends: $m_i = a + cZ_i$

Weekdays:

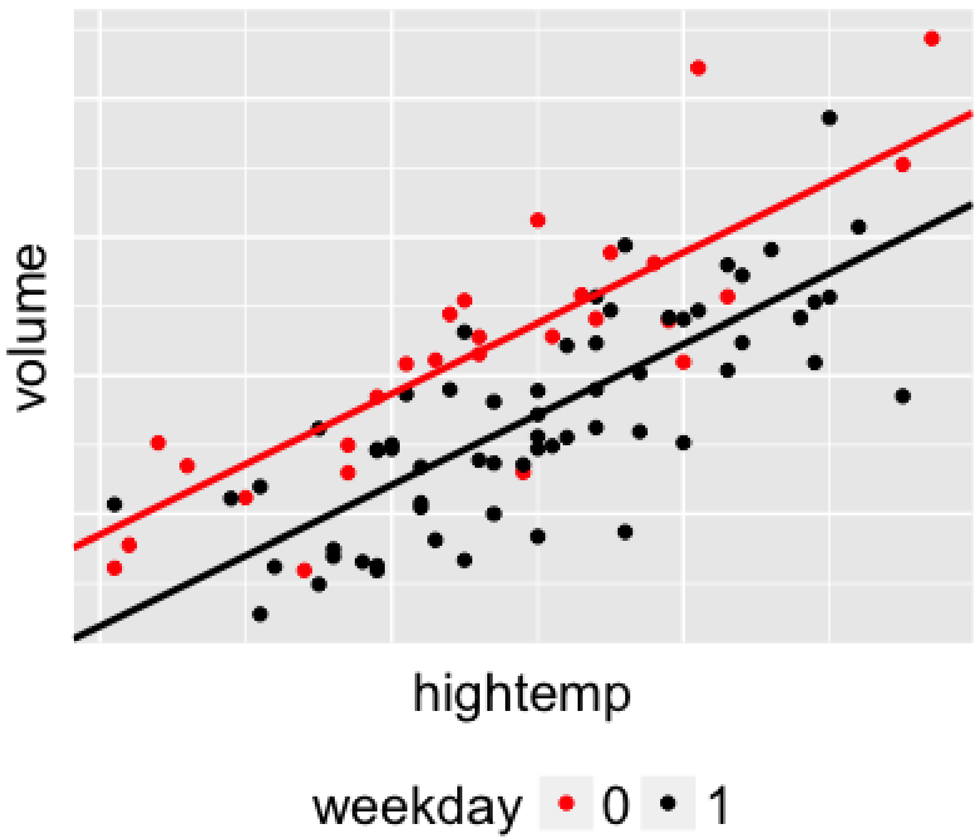
$$m_i = (a + b) + cZ_i$$

Modeling volume by temperature & weekday

Y_i = trail volume (# of users) on day i

X_i = 1 for weekdays, 0 for weekends

Z_i = high temperature on day i (in $^\circ$F)



$$Y_i \sim N(m_i, s^2)$$

$$m_i = a + bX_i + cZ_i$$

Weekends: $m_i = a + cZ_i$

Weekdays:
$$m_i = (a + b) + cZ_i$$

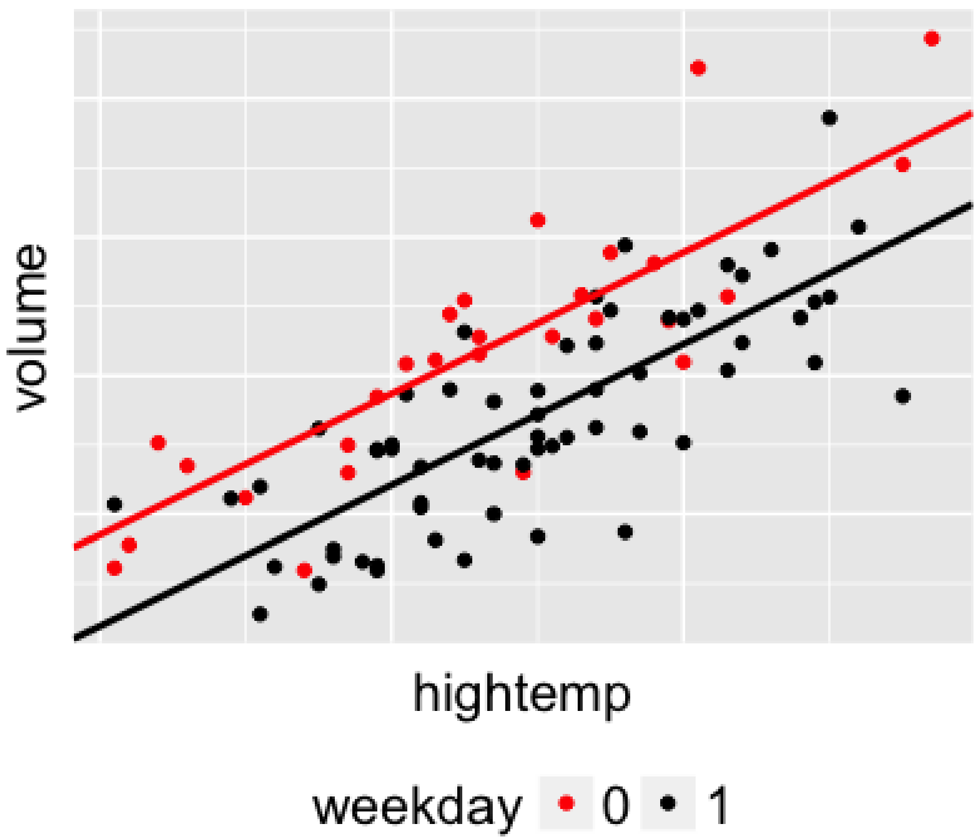
Modeling volume by temperature & weekday

$m_i = a + bX_i + cZ_i$

Weekends: $m_i = a + cZ_i$

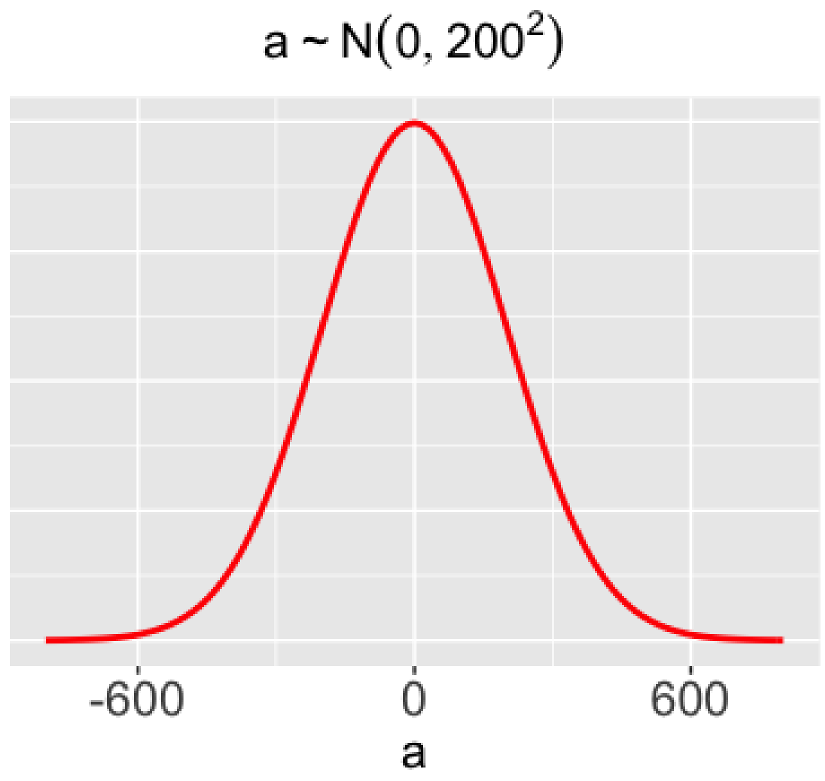
Weekdays:
$m_i = (a + b) + cZ_i$

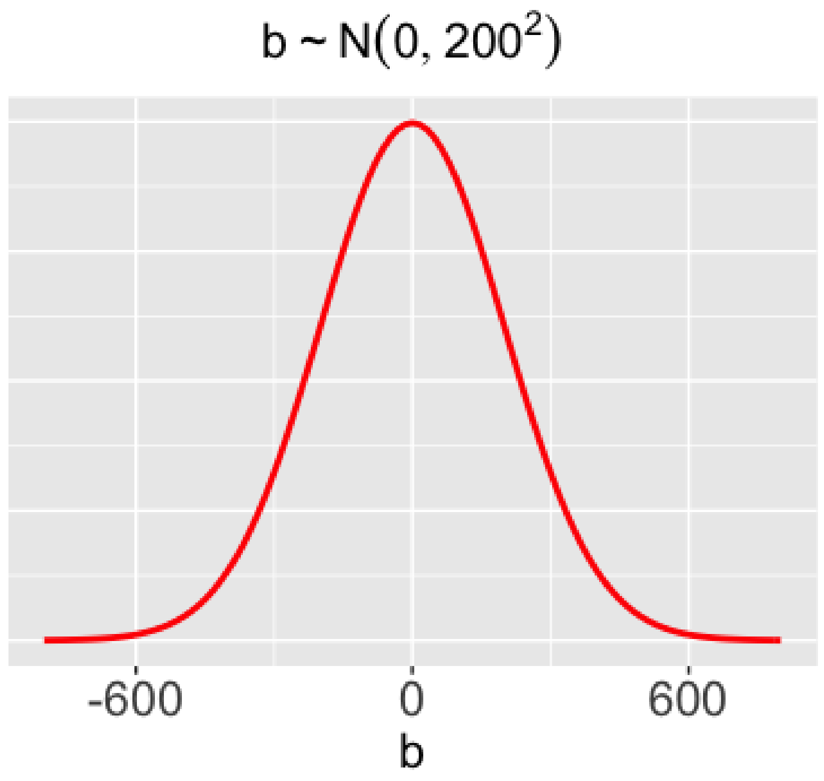
- a = weekend y-intercept
- $a + b$ = weekday y-int.
- b = contrast between weekday vs weekend y-intercepts



- c = common slope
- s = residual standard deviation

Priors for a and b

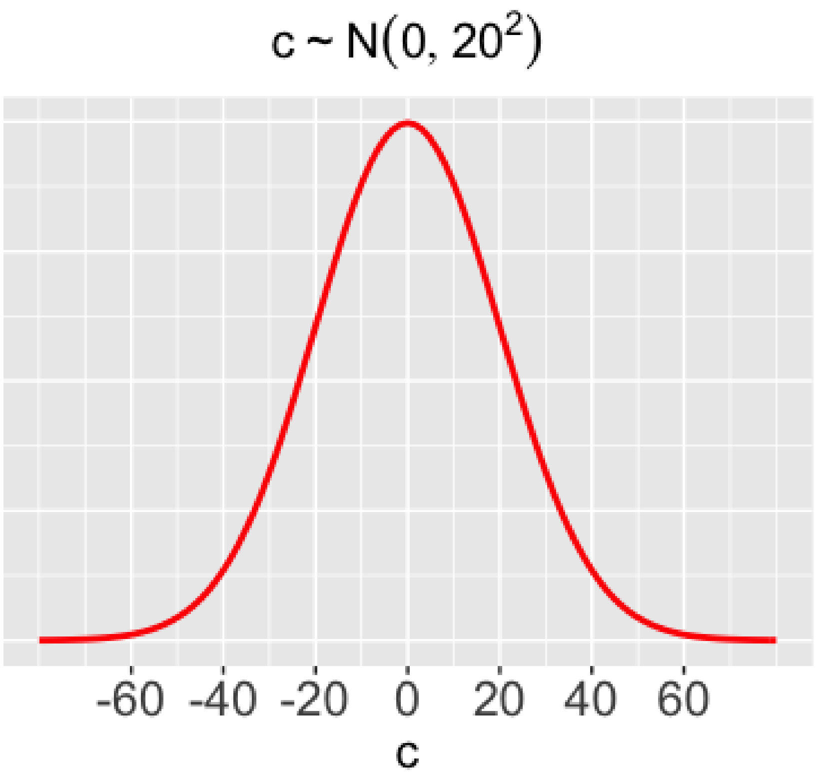


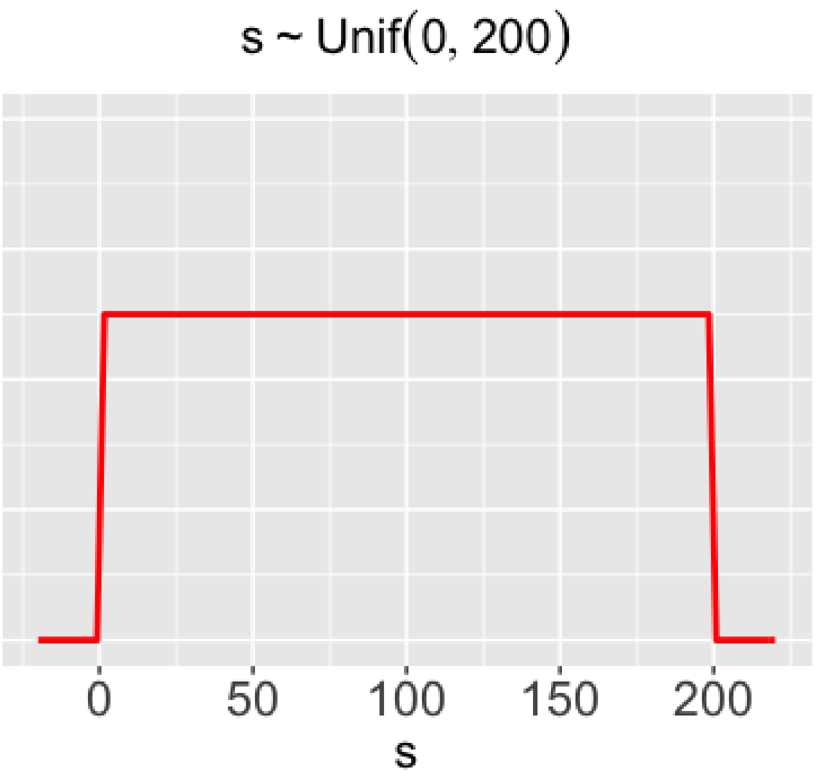


We lack certainty about the y-intercept for the relationship between temperature & *weekend* volume.

We lack certainty about how typical volume compares on weekdays vs weekends of similar temperature.

Priors for c and s



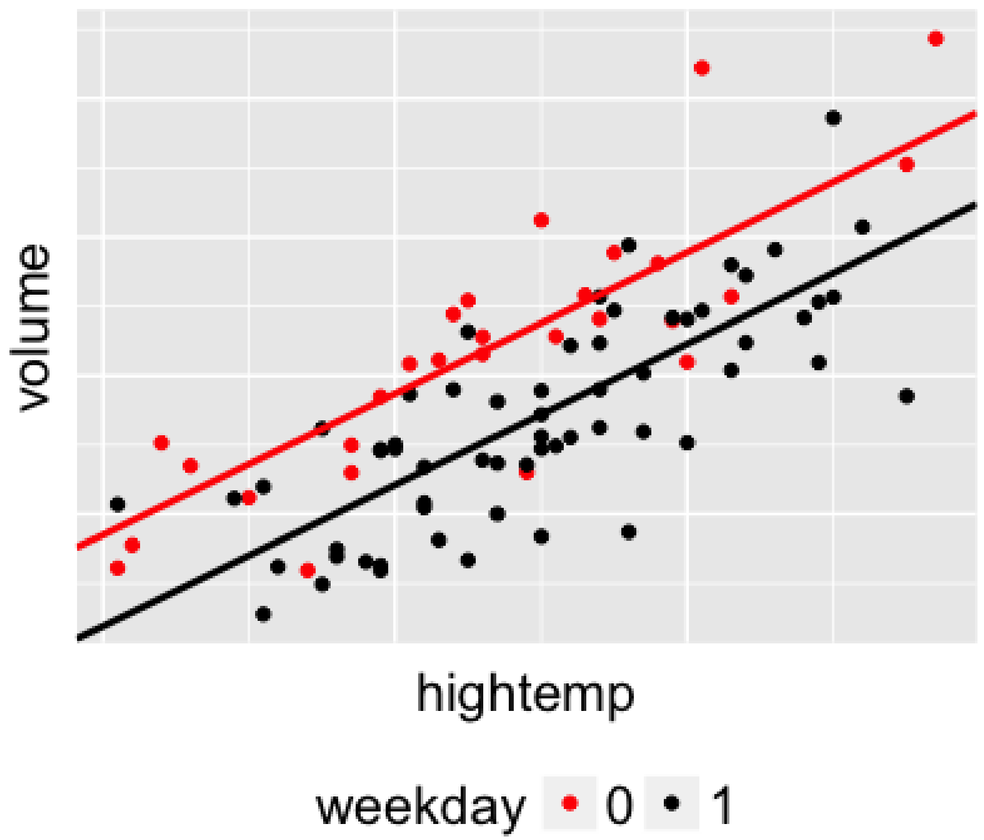


Whether on weekdays or weekends, we lack certainty about the association between trail volume & temperature.

The typical deviation from the trend is equally likely to be anywhere between 0 and 200 users.

Bayesian model of volume by weekday status

$$Y_i \sim N(m_i, s^2)$$
$$m_i = a + bX_i + cZ_i$$
$$a \sim N(0, 200^2)$$
$$b \sim N(0, 200^2)$$
$$c \sim N(0, 20^2)$$
$$s \sim \text{Unif}(0, 200)$$



DEFINE the Bayesian model in RJAGS

$$Y_i \sim N(m_i, s^2)$$
$$m_i = a + bX_i + cZ_i$$
$$a \sim N(0, 200^2)$$
$$b \sim N(0, 200^2)$$
$$c \sim N(0, 20^2)$$
$$s \sim \text{Unif}(0, 200)$$

```
rail_model_2 <- "model{
  # Likelihood model for Y[i]
  for(i in 1:length(Y)) {
    Y[i] ~ dnorm(m[i], s^(-2))
    m[i] <- a + b[X[i]] + c * Z[i]
  }

  # Prior models for a, b, c, s
  a ~ dnorm(0, 200^(-2))
  b[1] <- 0
  b[2] ~ dnorm(0, 200^(-2))
  c ~ dnorm(0, 20^(-2))
  s ~ dunif(0, 200)
}"
```

Let's practice!

BAYESIAN MODELING WITH RJAGS

Poisson regression

BAYESIAN MODELING WITH RJAGS

Alicia Johnson
Associate Professor, Macalester College

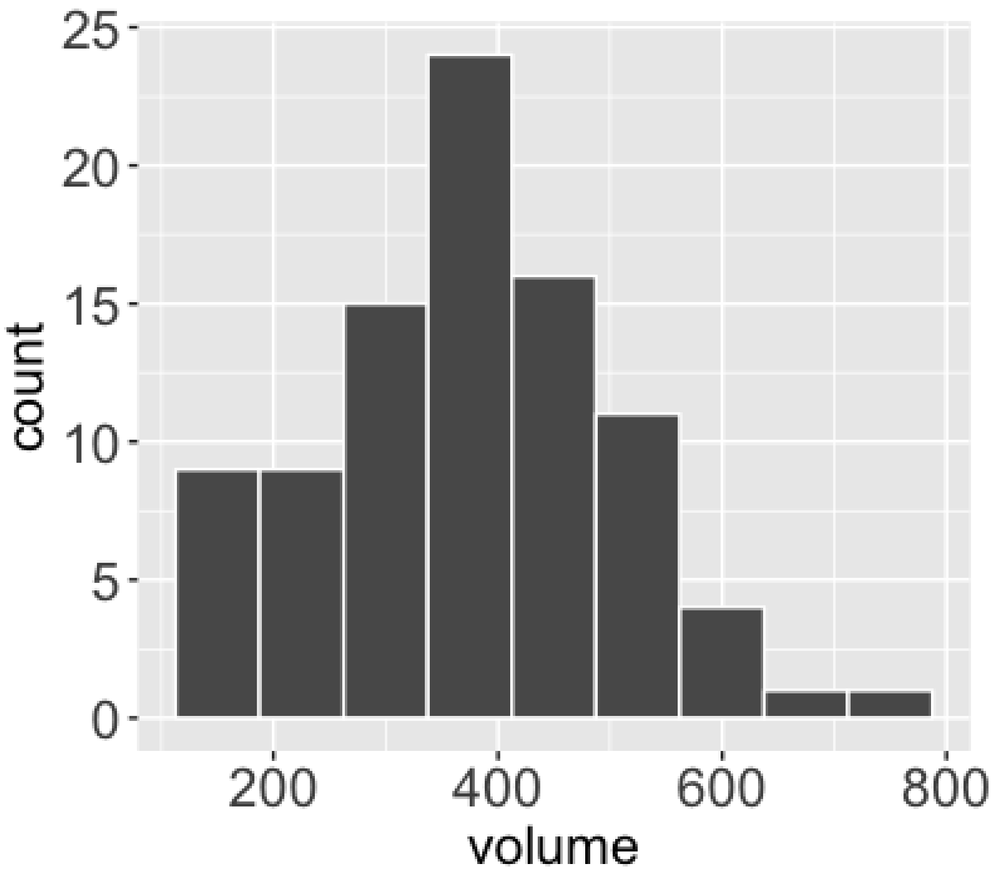
Normal likelihood structure

Y = volume (# of users) on a given day

$Y \sim N(m, s^2)$

Technically...

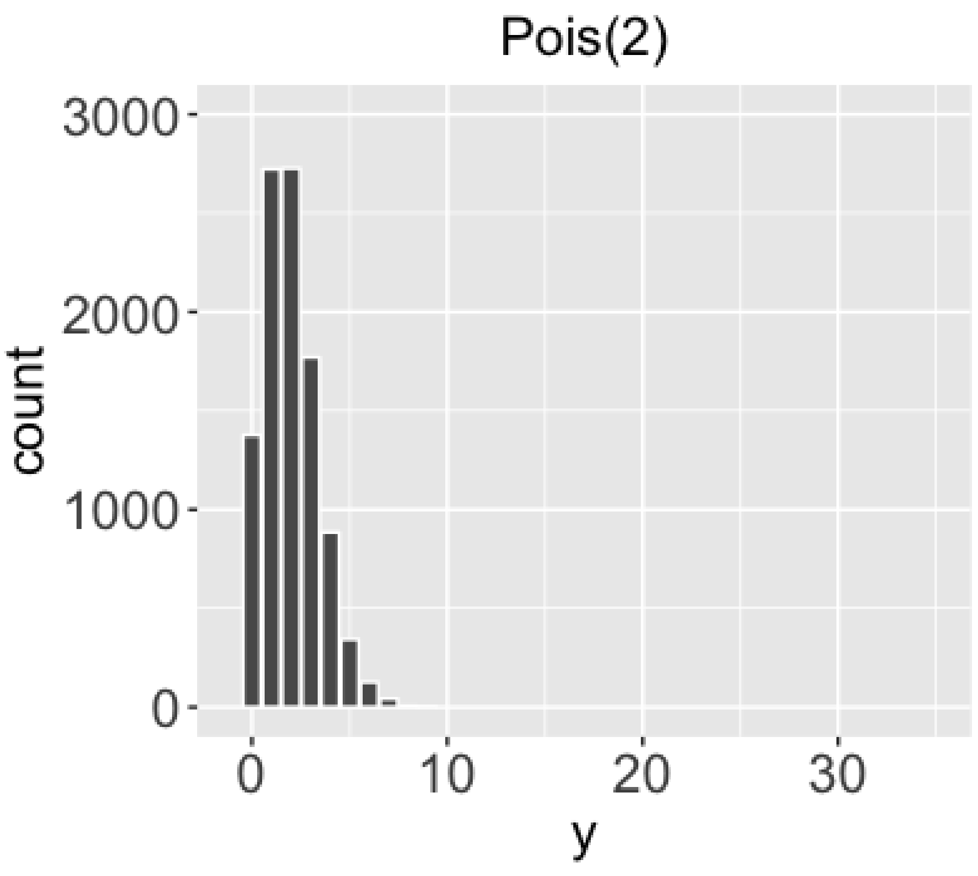
- The Normal model assumes Y has a continuous scale and can be negative.
- But Y is a discrete count and cannot be negative.



The Poisson model

Y = volume (# of users) on a given day
$Y \sim \mathrm{Pois}(l)$

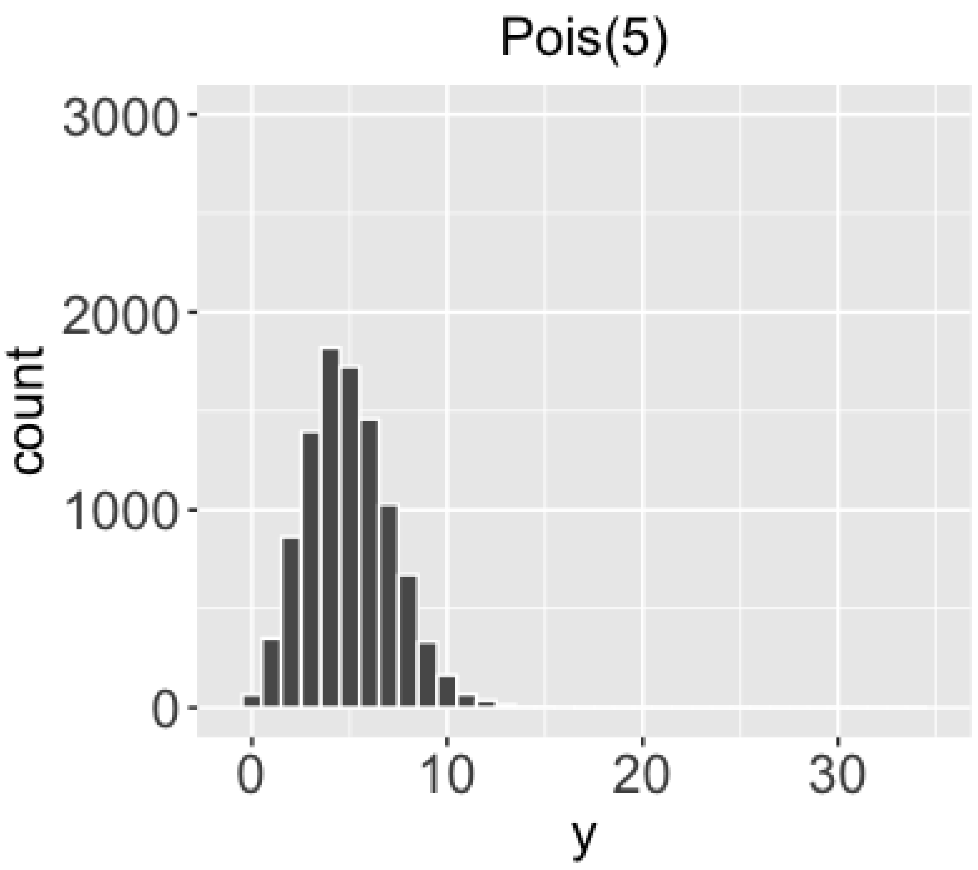
- Y is the # of independent events that occur in a fixed interval (0, 1, 2,...).
- *Rate parameter* l represents the typical # of events per time interval ($l > 0$).



The Poisson model

Y = volume (# of users) on a given day
$Y \sim \text{Pois}(l)$

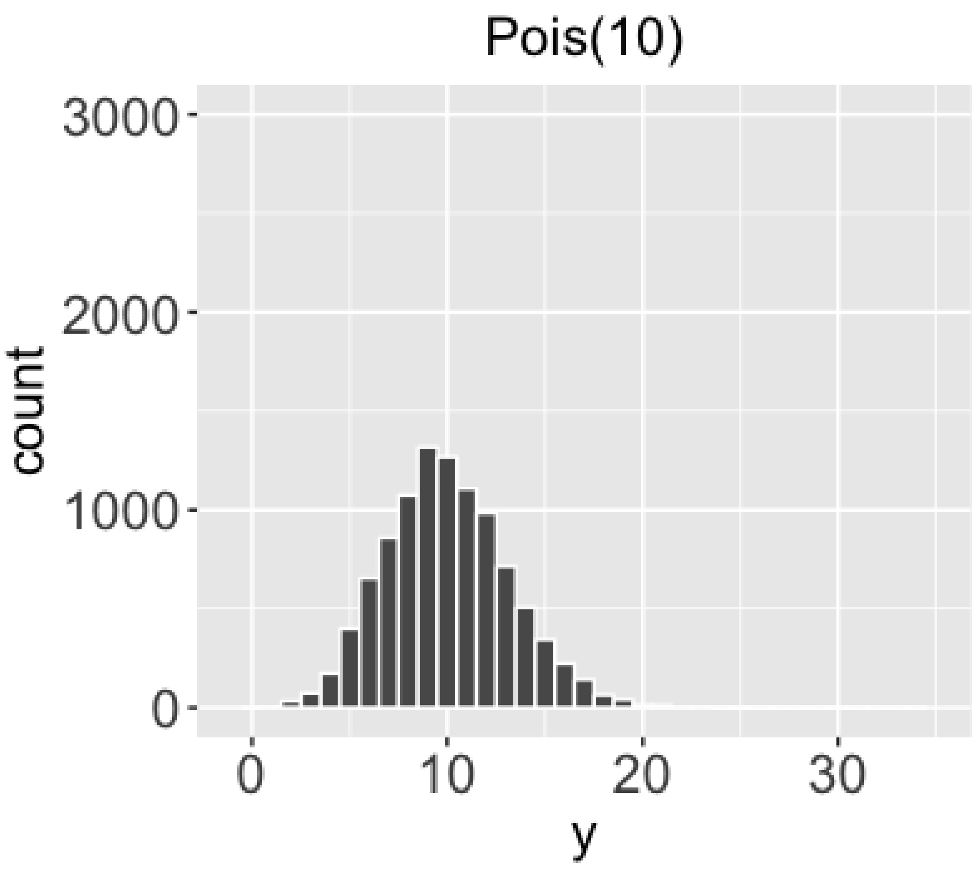
- Y is the # of independent events that occur in a fixed interval (0, 1, 2,...).
- *Rate parameter* l represents the typical # of events per time interval ($l > 0$).



The Poisson model

Y = volume (# of users) on a given day
$Y \sim \text{Pois}(l)$

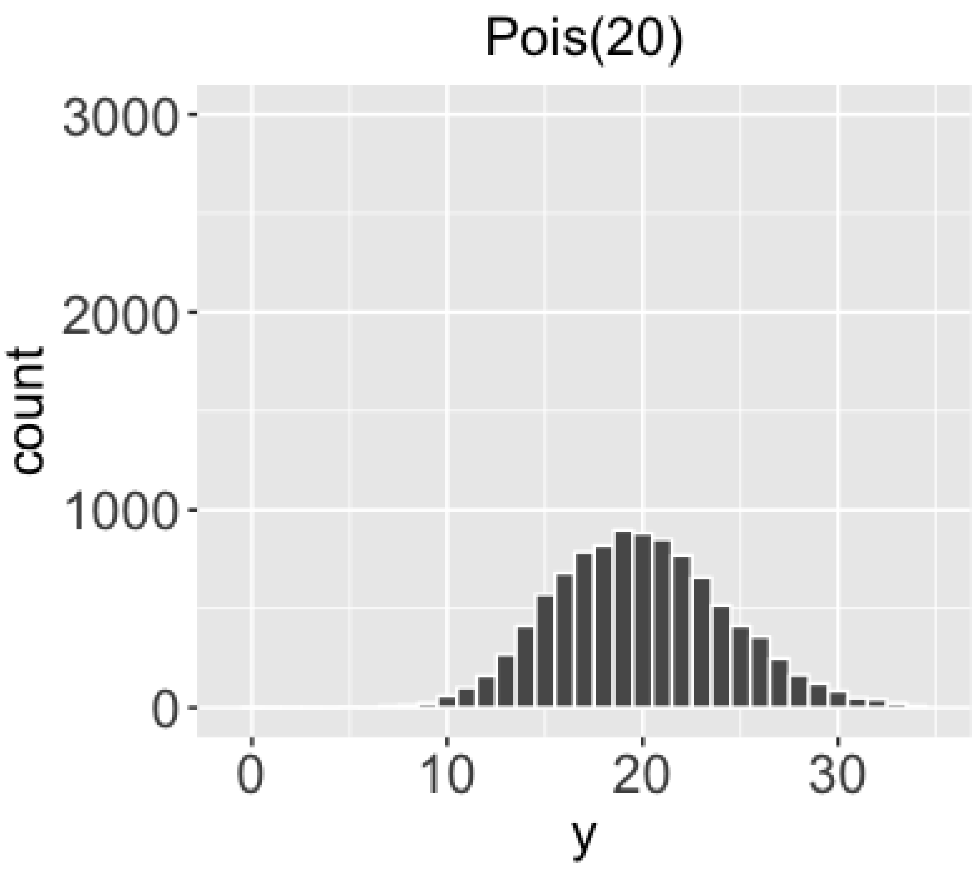
- Y is the # of independent events that occur in a fixed interval (0, 1, 2,...).
- *Rate parameter* l represents the typical # of events per time interval ($l > 0$).



The Poisson model

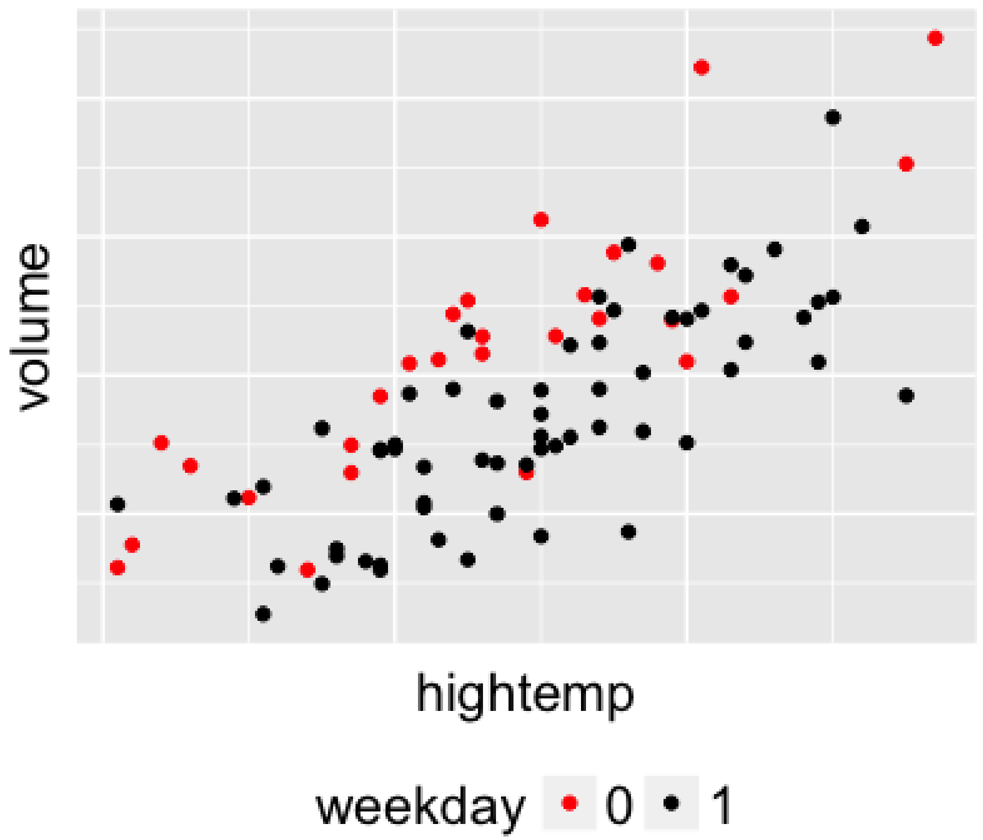
Y = volume (# of users) on a given day
$Y \sim \mathrm{Pois}(l)$

- Y is the # of independent events that occur in a fixed interval (0, 1, 2,...).
- *Rate parameter* l represents the typical # of events per time interval ($l > 0$).



Poisson regression

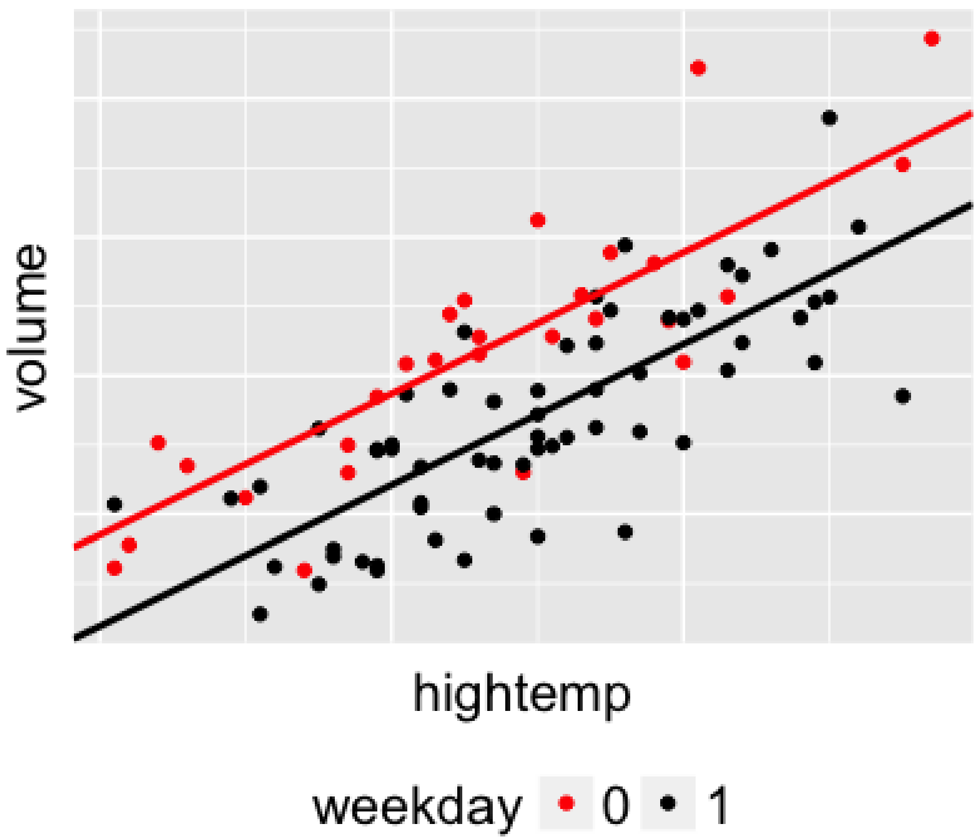
$Y_i \sim \text{Pois}(l_i)$ where $l_i > 0$



Poisson regression

$Y_i \sim \text{Pois}(l_i)$ where $l_i > 0$

$l_i = a + bX_i + cZ_i$



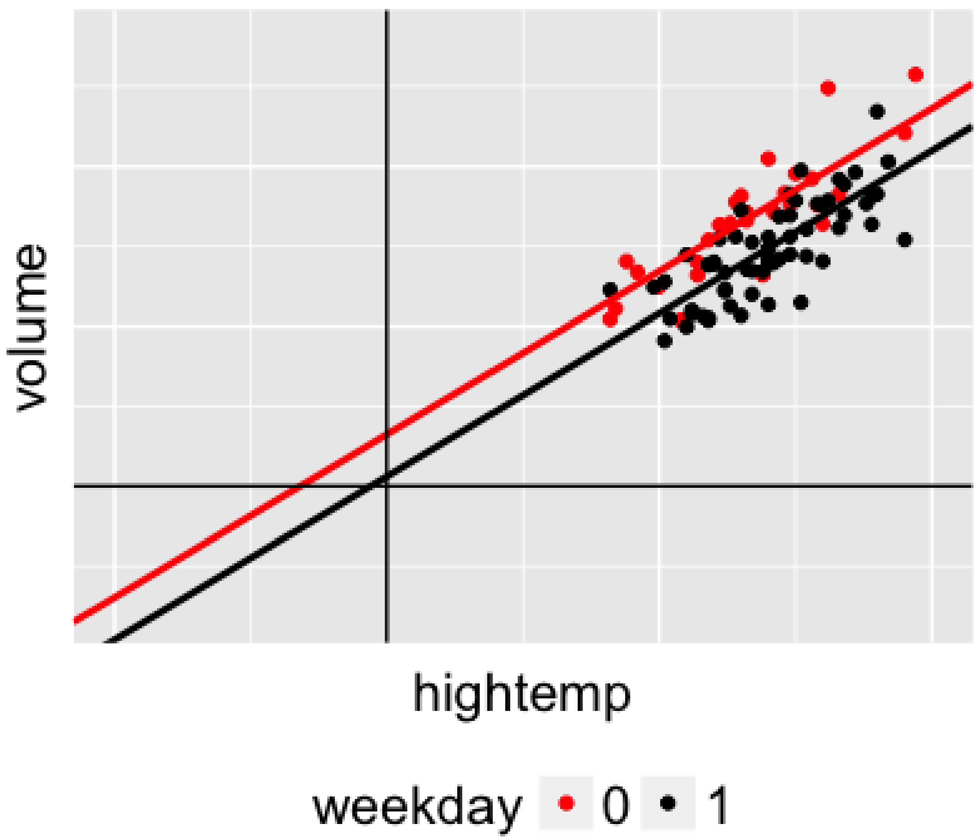
Poisson regression

$Y_i \sim \text{Pois}(l_i)$ where $l_i > 0$

$$l_i = a + bX_i + cZ_i$$

A problem:

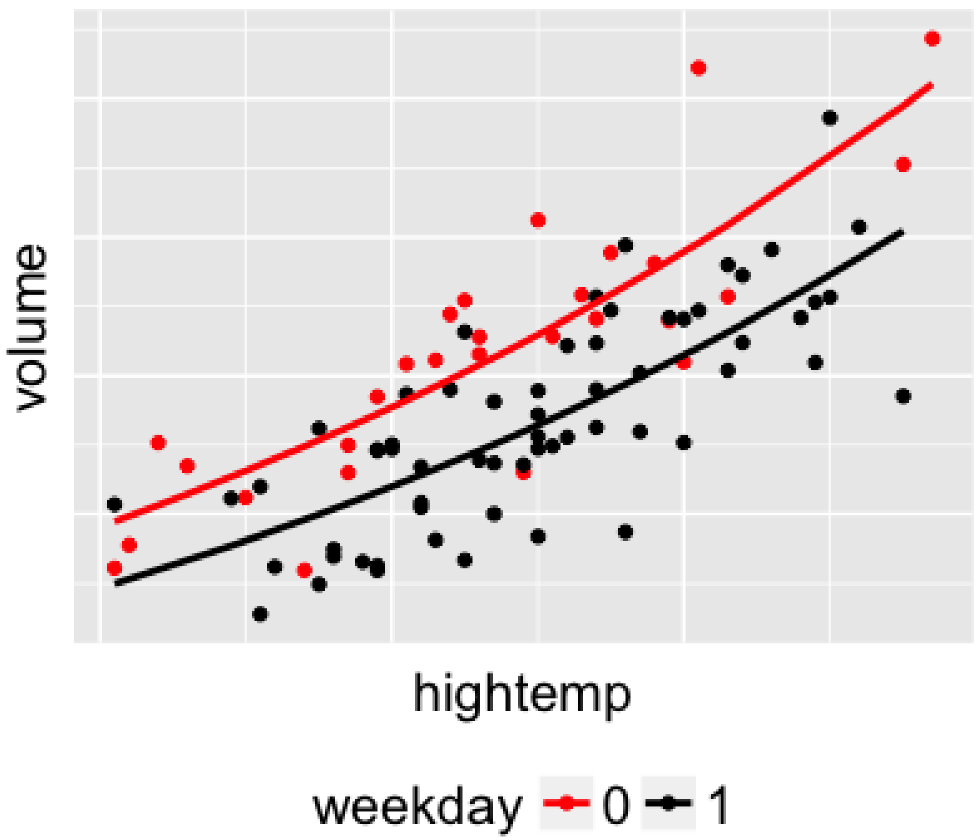
Linking l_i directly to the linear model assumes l_i can be negative.



Poisson regression

$Y_i \sim \text{Pois}(l_i)$ where $l_i > 0$

$log(l_i) = a + bX_i + cZ_i$



A solution:

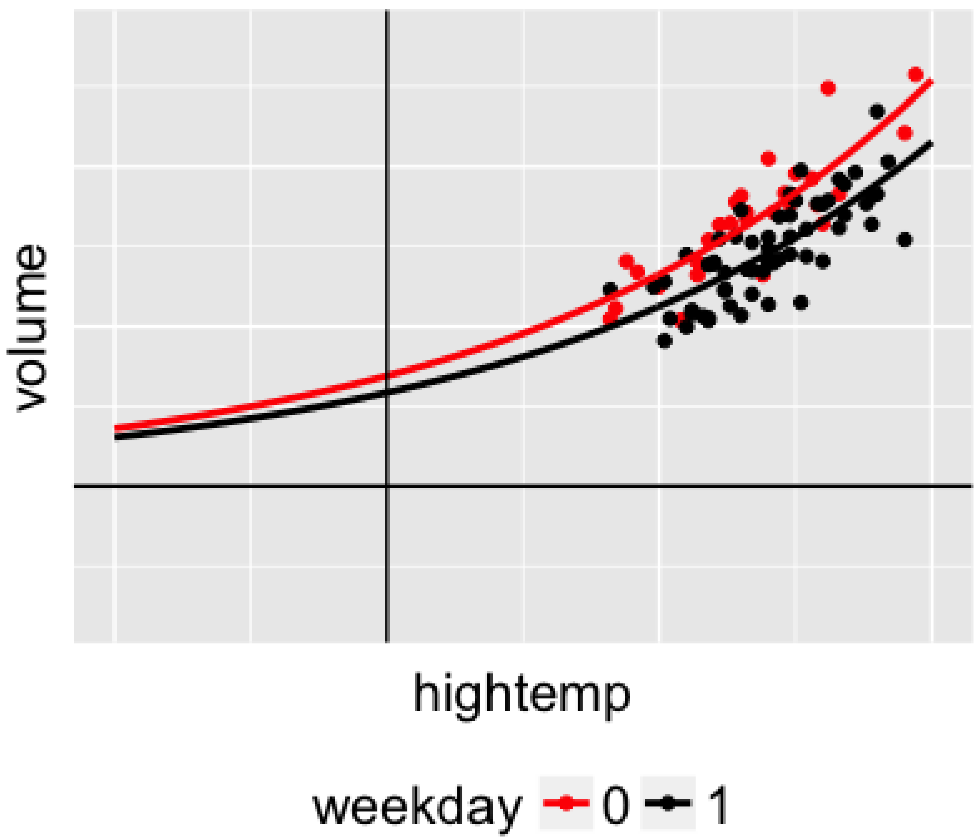
Use a log **link function** to link l_i to the linear model. In turn:

$$l_i = e^{a+bX_i+cZ_i}$$

Poisson regression

$Y_i \sim \text{Pois}(l_i)$ where $l_i > 0$

$$log(l_i) = a + bX_i + cZ_i$$



A solution:

Use a log **link function** to link l_i to the linear model. In turn:

$$l_i = e^{a+bX_i+cZ_i}$$

Poisson regression in RJAGS

$Y_i \sim \text{Pois}(l_i)$

$log(l_i) = a + bX_i + cZ_i$

$a \sim N(0, 200^2)$

$b \sim N(0, 2^2)$

$c \sim N(0, 2^2)$

```
poisson_model <- "model{
  # Likelihood model for Y[i]




  # Prior models for a, b, c




}"
```

Poisson regression in RJAGS

$Y_i \sim \text{Pois}(l_i)$

$log(l_i) = a + bX_i + cZ_i$

$a \sim N(0, 200^2)$

$b \sim N(0, 2^2)$

$c \sim N(0, 2^2)$

```
poisson_model <- "model{
  # Likelihood model for Y[i]





  # Prior models for a, b, c
  a ~ dnorm(0, 200^(-2))
  b[1] <- 0
  b[2] ~ dnorm(0, 2^(-2))
  c ~ dnorm(0, 2^(-2))
}"
```

Poisson regression in RJAGS

$Y_i \sim \text{Pois}(l_i)$

$log(l_i) = a + bX_i + cZ_i$

$a \sim N(0, 200^2)$

$b \sim N(0, 2^2)$

$c \sim N(0, 2^2)$

```
poisson_model <- "model{
  # Likelihood model for Y[i]
  for(i in 1:length(Y)) {
   Y[i] ~ dpois(l[i])

  }

  # Prior models for a, b, c
  a ~ dnorm(0, 200^(-2))
  b[1] <- 0
  b[2] ~ dnorm(0, 2^(-2))
  c ~ dnorm(0, 2^(-2))
}"
```

Poisson regression in RJAGS

$Y_i \sim \text{Pois}(l_i)$

$log(l_i) = a + bX_i + cZ_i$

$a \sim N(0, 200^2)$

$b \sim N(0, 2^2)$

$c \sim N(0, 2^2)$

```
poisson_model <- "model{
  # Likelihood model for Y[i]
  for(i in 1:length(Y)) {
   Y[i] ~ dpois(l[i])
   log(l[i]) <- a + b[X[i]] + c*Z[i]
  }

  # Prior models for a, b, c
  a ~ dnorm(0, 200^(-2))
  b[1] <- 0
  b[2] ~ dnorm(0, 2^(-2))
  c ~ dnorm(0, 2^(-2))
}"
```

Caveats

$Y \sim \text{Pois}(l_i)$

- Assumption: Among days with similar temperatures and weekday status, variance in Y_i is equal to the mean of Y_i.
- Our data demonstrate potential **overdispersion** - the variance is larger than the mean.
- Though not perfect, this model is an OK place to start.

Let's practice!

BAYESIAN MODELING WITH RJAGS

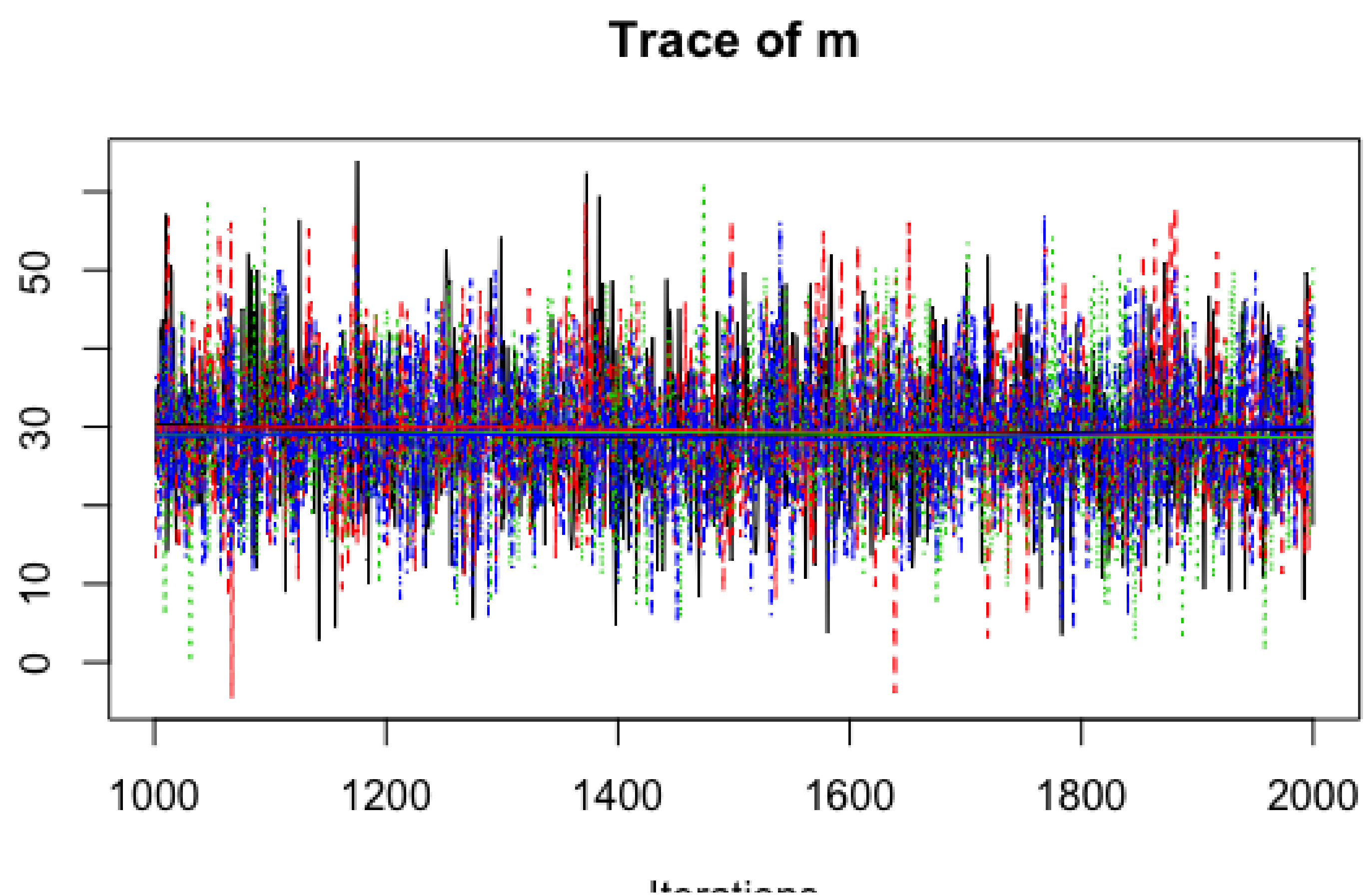
Conclusion

BAYESIAN MODELING WITH RJAGS

Alicia Johnson
Associate Professor, Macalester College

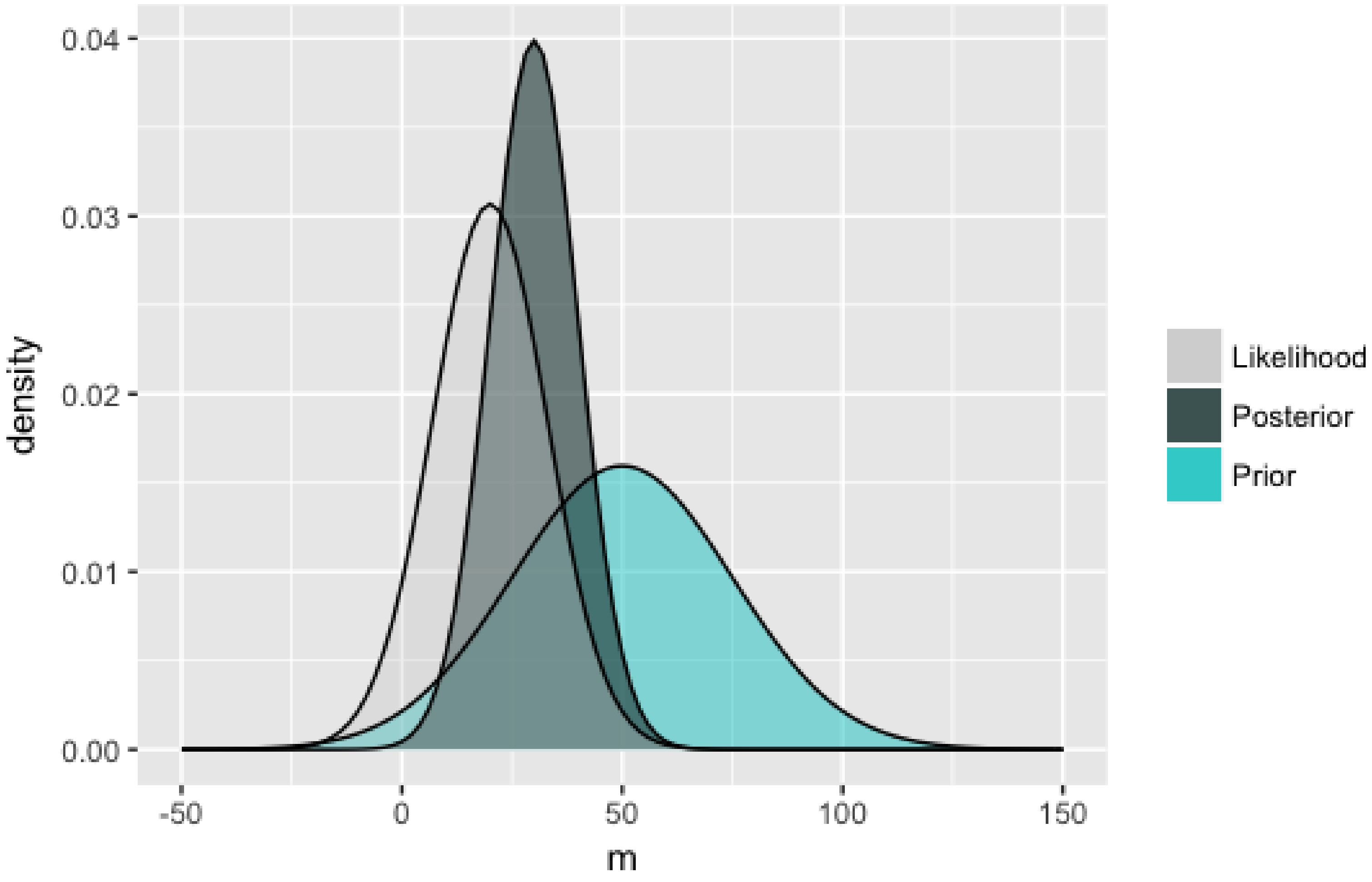
Bayesian modeling with RJAGS

- Define, compile, & simulate intractable Bayesian models.
- Explore the Markov chain mechanics behind RJAGS simulation.



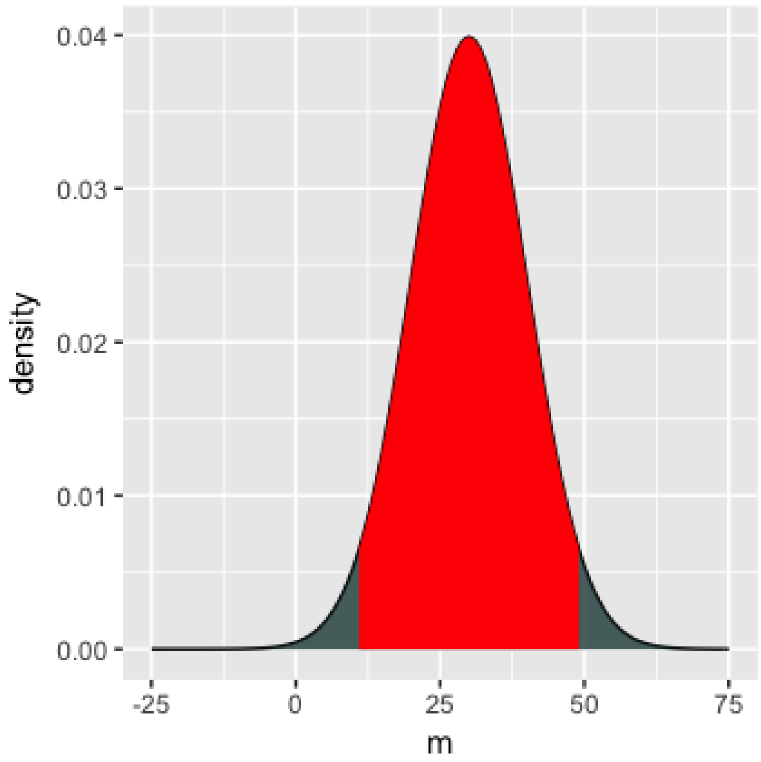
The power of Bayesian modeling

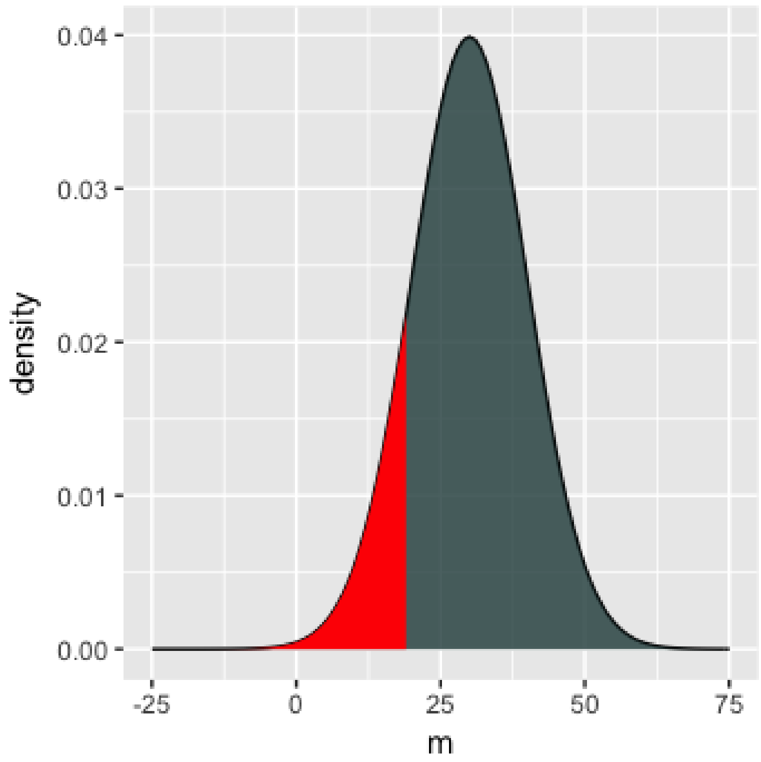
- Combine insights from your data *and* priors to inform posterior insights.



The power of Bayesian modeling

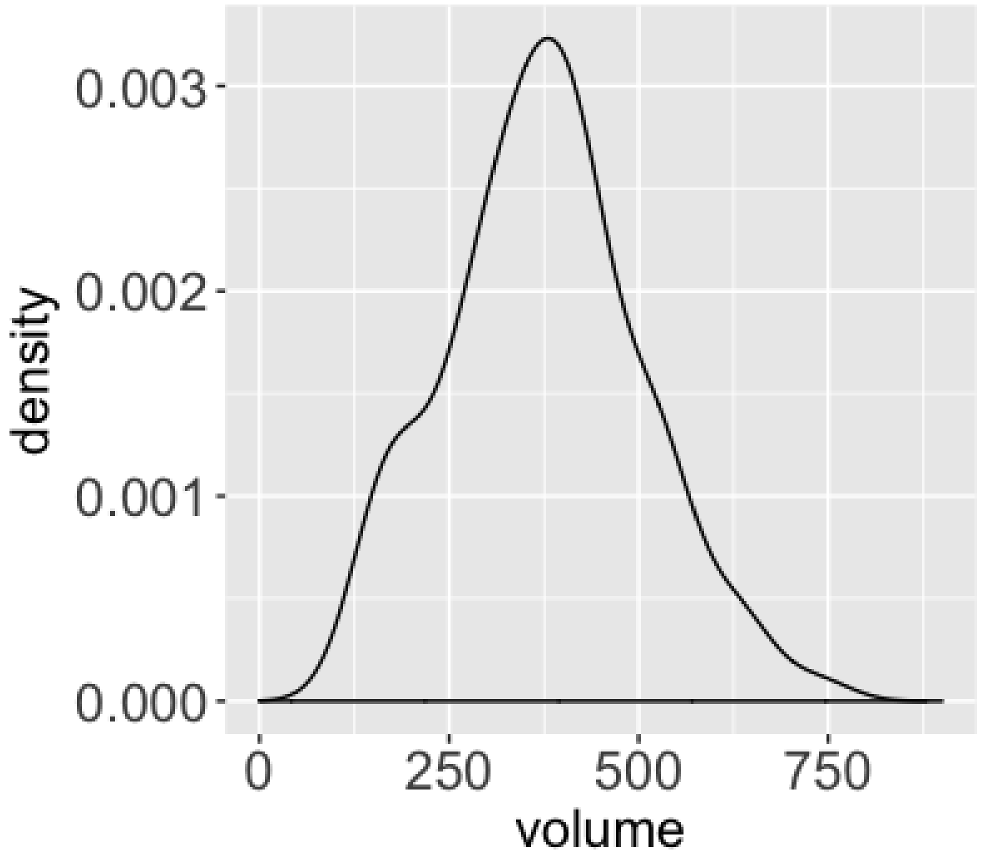
- Combine insights from your data *and* priors to inform posterior insights.
- Conduct intuitive posterior inference: posterior credible intervals & probabilities.





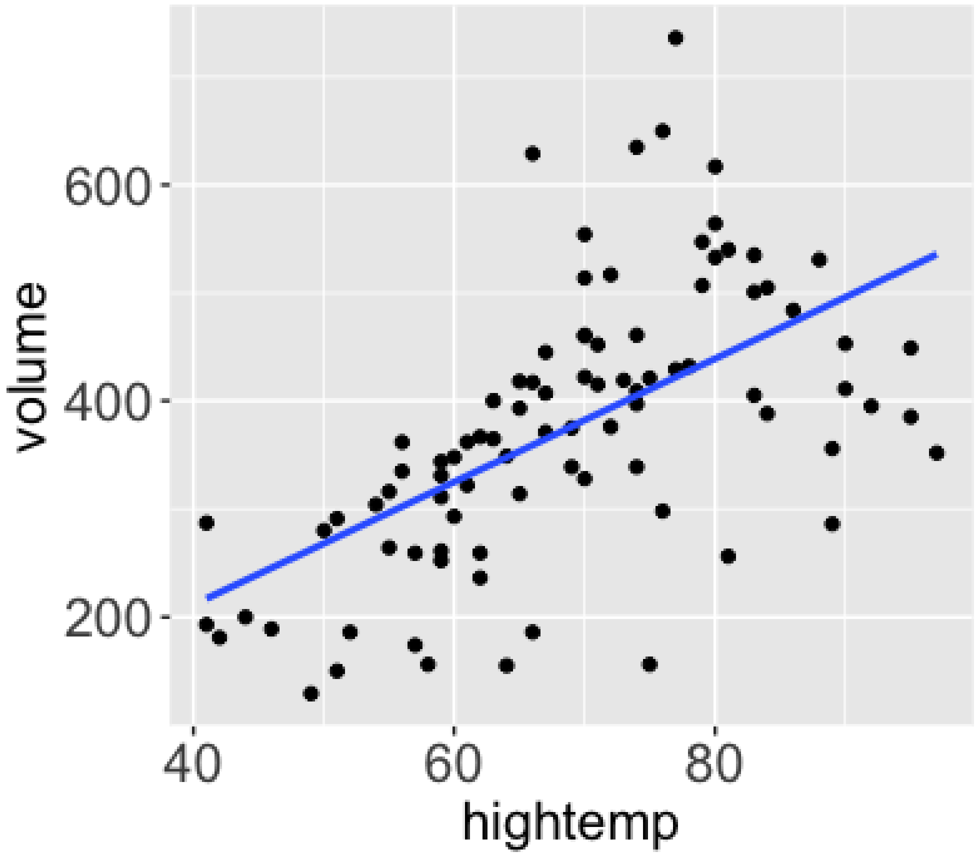
Foundational, flexible, & generalizable Bayesian models

```
my_model <- "model{
  # Likelihood model
  for(i in 1:length(Y)) {
    Y[i] ~ dnorm(m, s^(-2))
  }

  # Prior models
  m ~ dnorm(...)
  s ~ dunif(...)
}"
```



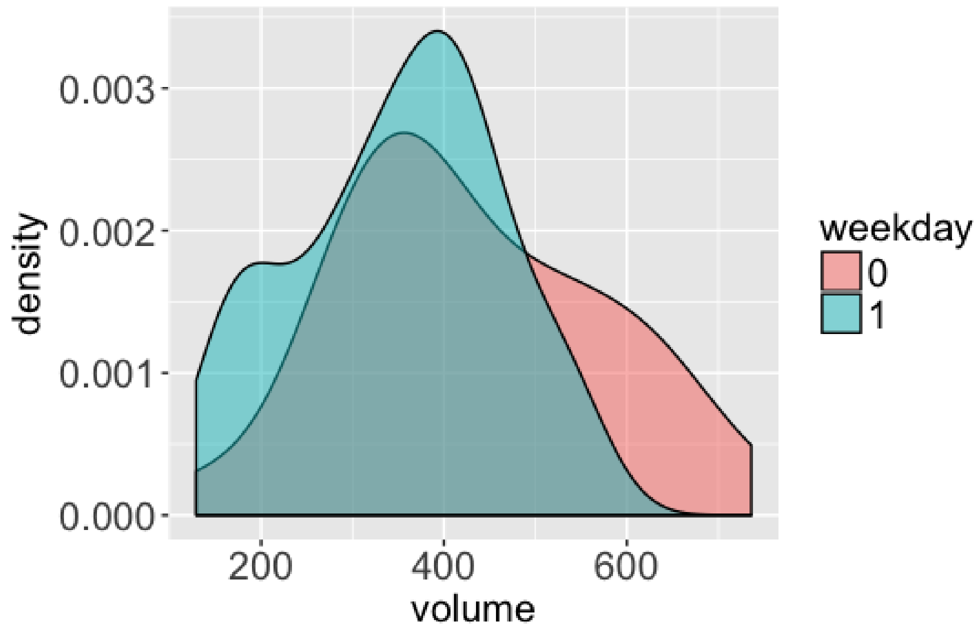
Foundational, flexible, & generalizable Bayesian models

```
my_model <- "model{
  # Likelihood model
  for(i in 1:length(Y)) {
    Y[i] ~ dnorm(m[i], s^(-2))
    m[i] <- a + b * X[i]
  }

  # Prior models
  a ~ dnorm(...)
  b ~ dnorm(...)
  s ~ dunif(...)
}"
```



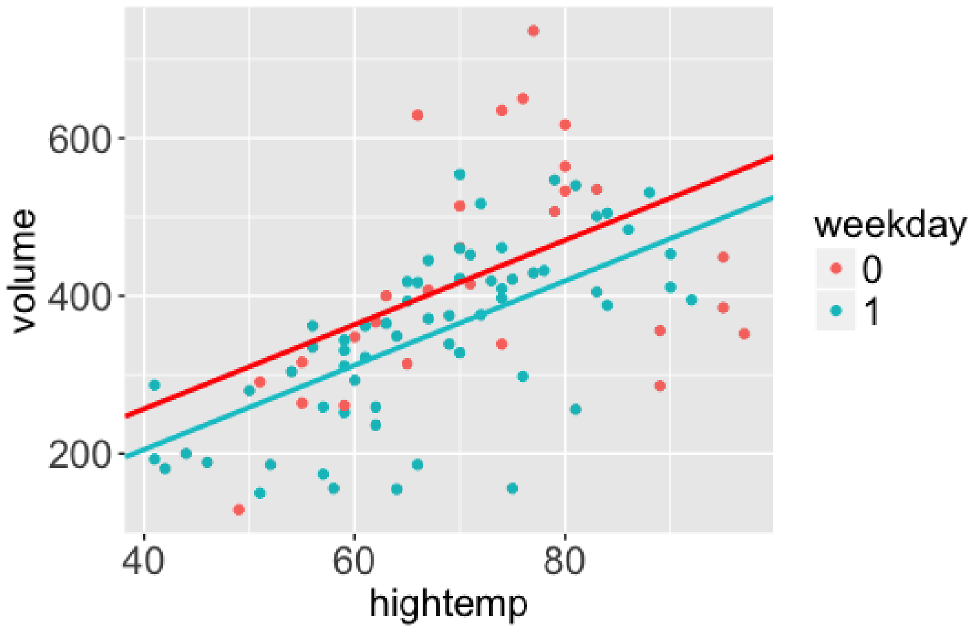
Foundational, flexible, & generalizable Bayesian models

```
my_model <- "model{
  # Likelihood model
  for(i in 1:length(Y)) {
    Y[i] ~ dnorm(m[i], s^(-2))
    m[i] <- a + b[X[i]]
  }

  # Prior models
  a ~ dnorm(...)
  b[1] <- 0
  b[2] ~ dnorm(...)
  s ~ dunif(...)
}"
```



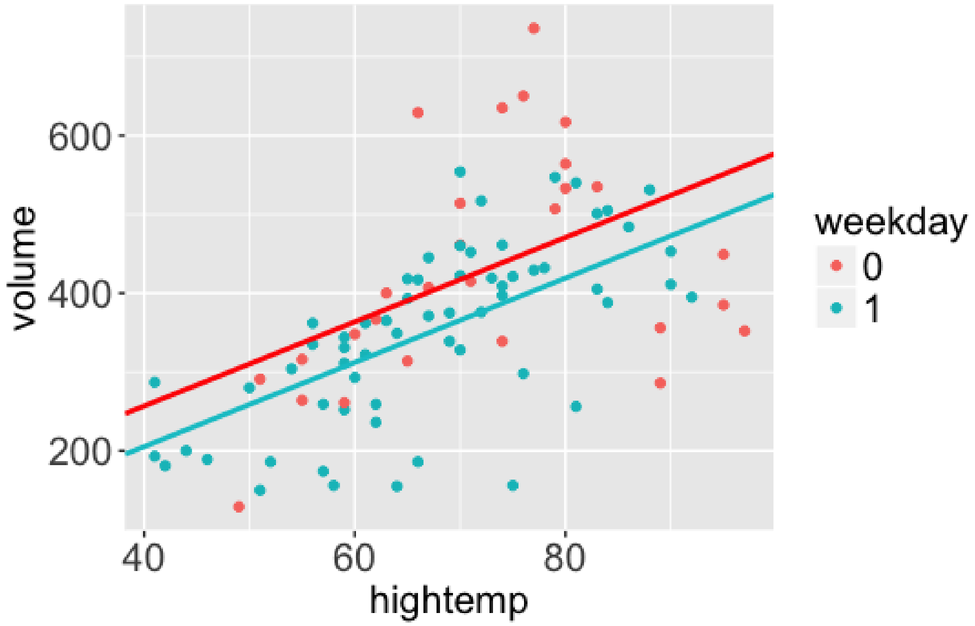
Foundational, flexible, & generalizable Bayesian models

```
my_model <- "model{
  # Likelihood model
  for(i in 1:length(Y)) {
    Y[i] ~ dnorm(m[i], s^(-2))
    m[i] <- a + b[X[i]] + c * Z[i]
  }
  # Prior models
  a ~ dnorm(...)
  b[1] <- 0
  b[2] ~ dnorm(...)
  c ~ dnorm(...)
  s ~ dunif(...)
}"
```



Foundational, flexible, & generalizable Bayesian models

```
my_model <- "model{
 # Likelihood model
 for(i in 1:length(Y)) {
 Y[i] ~ dpois(l[i])
 log(l[i]) <- a + b[X[i]] + c*Z[i]
 }
 # Prior models
 a ~ dnorm(...)
 b[1] <- 0
 b[2] ~ dnorm(...)
 c ~ dnorm(...)
}"
```



Thank you!

BAYESIAN MODELING WITH RJAGS